Detecting News Topics from Microblogs using Sequential Pattern Mining

a thesis submitted to
the science and engineering faculty
of Queensland University of Technology
in fulfilment of the requirements for the degree of
Doctor of Philosophy

Cher Han Lau

Science and Engineering Faculty
Queensland University of Technology

March 2014

Copyright in Relation to This Thesis

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature:

Date: 3rd March 2014

To my supervisors, family members and friends

Abstract

Large amounts of news-related information are available on microblog platforms such as Twitter, but this information is scattered and the volume of tweets makes it both difficult and tedious for humans to filter irrelevant information and to extract meaningful topics. Traditional topic and event detection techniques are mostly developed using term-based models and are limited to specific domains or events. Terms-based techniques are not suitable for Twitter as they are sensitive to noise and require additional effort for interpreting the results. Term-based topics are represented using Bag-of-Words (BOW) without retaining term relationships, which makes the topics harder to understand compared with pattern-based topics. This motivates the needs for a microblog news detection framework.

This thesis presents a microblog news detection framework using a sequential pattern model. We design Pattern Model for Microblog (PMM) to represent topics as an ordered list of terms. PMM effectively captures key topics in news, such as persons, locations, organizations and events. Topic importance is then measured by evaluating topic weights using pattern properties and Twitter characteristics.

The main output of this research is an automatic news detection framework that works across multiple domains and is not limited to specific topic types. The research contributes to topic identification in microblogs using pattern-based model, and improves news topics identification using multiple metrics. Experiments with a large-scale standard public dataset (16 million tweets) show that our framework is effective in detecting news topics and can be applied into various applications such as editorial support systems to help news editors in finding potential news topics, or to assist authorities in monitoring developments during natural disasters and critical situations.

Keywords

microblogs, sequential pattern mining, news topic detection, text mining

Acknowledgments

It has been a long journey since I began my doctoral program. Throughout the journey, I received lots of support from my supervisors, colleagues, friends and family, all of whom I would like to thank.

My sincerest thanks goes to my supervisor, Associate Professor Dian Tjondronegoro, for his self-less contribution of time and guidance, not only as a supervisor, but also a mentor and close personal friend. I also thank my associate supervisors: Professor Yuefeng Li for his support and advice, and Associate Professor Yue Xu for her valuable opinions.

I owe gratitude to Smartservices CRC, for the support and training throughout my candidature. Special mention goes to Fairfax Digital for their inputs during my study and the precious opportunity for a three-week internship at their Sydney Headquarters. I would also like to thank the HDR support team in SEF, and the student society for their administrative help and enabling social engagement with other peers. Specifically I would like to thank Jack Tseng, Daniel Tao, Wei Song, Johannes Sasongko and Ligang Zhang for their greatly appreciated support, understanding and friendship throughout my PhD journey.

Last but not least, words cannot express my thanks to my parents and fiancée for their love, encouragement, and support throughout my study period.

Cher Han Lau
3^{rd} March 2014

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Microblog services such as Twitter are becoming popular as the main channel where users spread information and express opinions during events. These events cover everything from entertainment, sport and politics to local events. Twitter is used heavily to spread information during these events; sometimes, tweets (Twitter microblog posts) even travel faster and break news before media outlets can do this (e.g. death of Michael Jackson[1]). Organizations like NASA even choose to release official announcements through Twitter first (Figure 1.1), which has caught media outlets off guard[2]. This amount of news information in microblogs motivates the needs for a news topics detection system.

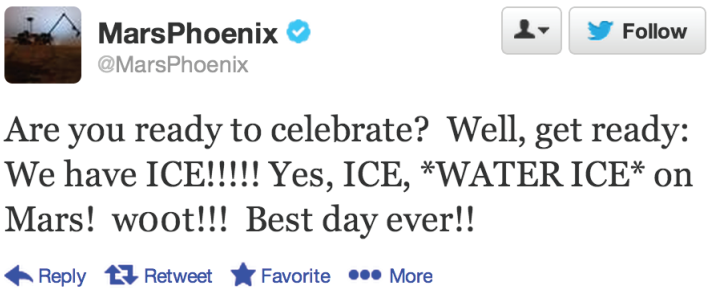


Figure 1.1: NASA announce evidence of water on Mars through Twitter

[1] http://www.dailymail.co.uk/sciencetech/article-1195651/How-Michael-Jacksons-death-shut-Twitter-overwhelmed-Google–killed-Jeff-Goldblum.html

[2] http://www.wired.com/wiredscience/2008/06/mars-phoenix-tw/

Finding news topics from microblogs is not easy. Twitter delivers over 400 million tweets daily[3], which makes it almost impossible for humans to skim and filter irrelevant tweets to find news topics. The word limit of tweets further complicates the problem by causing users to apply abbreviations, slang and emoticons to compress information. This has resulted in a noisy vocabulary. Furthermore, sharing urls and using special Twitter syntaxes such as retweet, mention and hashtag increase the complexity required to process and understand tweets. Each tweet contributes to only a smaller fragment of a big topic, which makes it difficult to obtain sensible topic clusters.

Previous work on detecting news from microblogs is limited to specific applications with different content focus. *TwitterStand* (Sankaranarayanan et al., 2009) focuses on detecting news topics around specific geographical area, *Submlr* (Shou et al., 2013) summarizes tweet streams and Petrovic et al. (2010) detect the first instance of a story posted to Twitter. Other systems such as Twitinfo (Marcus et al., 2011) focuses on helping users to visualize key moments during an event and Eddi (Bernstein et al., 2010) focuses on topic browsing, but limited to single user stream. Other related news detection techniques focus on particular type of events such as elections (Diakopoulos and Shamma, 2010; Shamma et al., 2009), natural disasters (Bruns et al., 2012; Li and Rao, 2010; Sakaki et al., 2010), and local incidents such as fire (Abel et al., 2012) or factory strikes (Agarwal et al., 2012). There is still a lack of systems that focus on detecting news topics, by considering multiple information aspects.

One research area closely related to news detection is the Topic Detection and Tracking (TDT). TDT research has achieved great success on full length news articles, but these techniques have yet to be fully implemented on microblogs. One key challenge is the source quality: TDT sources are news articles that contain rich news topics and are well-structured, where microblog data is noisy and unstructured. Performance of TDT techniques relies on good quality input, but the quality of tweets varies as it contain not only news, but also chatter, conversations and question answering. Furthermore, text processing techniques are not directly applicable to short text such as tweet, as statistics

[3]http://www.theverge.com/2012/6/6/3069424/twitter-400-million-total-daily-tweets

cannot be easily derived (Metzler et al., 2007).

Success of a news topic detection system relies largely on its topic detection process. Many of the topic detection methods in microblogs use term-based methods such as TF-IDF (Efron, 2011; Lee et al., 2011a). Term-based methods are mature and easy to implement, but are sensitive to noise (Naveed et al., 2011b). Topics captured by term-based methods are usually presented as Bag-of-Words (BOW), in which a topic is represented as a collection of words without any ordering. This makes the topics harder to understand and requires additional effort to interpret. Previous studies from text mining field suggest that a pattern-based model such as Frequent Pattern Mining is capable of addressing the gaps in term-based approaches (Li et al., 2010; Wu et al., 2004; Kim et al., 2012), but these techniques have yet to be applied on microblogs.

This thesis presents a research work on a microblog news detection framework that aims to address the gaps above. The proposed framework presents a topic detection model based on pattern mining techniques, to capture topics using sequential patterns. Topics are then evaluated for their news relevance using context and content information, combined with other Twitter specific properties. The framework is able to detect news from microblogs without limiting to any specific event or news category.

1.2 Research Problems

Detecting news topics from microblogs is challenging due to the immense scale of volume and their short characteristic. Many unique features of microblogs have been exploited in event detection studies, and it is important to investigate how these features can be used for news detection. In order to detect news topics from microblogs effectively, the following questions need to be answered:-

- **How do process and extract features from microblogs?** Microblog is a completely different type of text compared with well-structured documents such as webpages, news articles and blogs. Extracting representative features from unstructured microblogs is the key for subsequent processing tasks.

- **How do we extract topics from microblogs using patterns:** Previous topic detection methods use statistical term-based techniques on long documents with rich content. But microblog is short and noisy which makes it difficult to gather contextual information and statistics for topic detection. Term-based topics are presented as collection of terms which are hard to understand. How do we extract topics from microblogs that are meaningful and be interpreted more easily?

- **How do we evaluate news relevance for extracted topics?** Microblogs contain strong opinionated information and possess unique characteristics. These features are previously studied for different events but their significance has not been shown using a single model for news detection. How can we utilize these features to evaluate news relevance for topics?

1.3 Aims and Objectives

This study covers the following aspects for detecting news from microblogs. We first investigate different microblogs properties: content, context and Twitter activities, to extract key features for processing. A topic detection algorithm is developed using data mining techniques to represent topics using sequential patterns instead of terms. Pattern representations allows us to capture more meaningful topics and to reduce noisy and redundant information.

We then present an algorithm to rank the discovered topics using different metrics to measure the probability of a topic to become a news topic. We consider temporal, public opinions and Twitters activity, to provide a more accurate news-relevance computation model.

1.4 Contributions and Significance

This research presents a novel contribution to the advancement of short-text modelling in text mining field, and news topics detection in microblog study. It improves the microblog

features using state-of-the-art text mining techniques, and show that patterns are suitable for use in microblogs and improve term-based representations.

For topic detection in microblogs, we develop a set of algorithms using pattern properties to reduce noises and remove redundant information. Pattern weights are distributed appropriately according to their indicative power and the amount of information carried.

This thesis also investigates various Twitter specific features, which in previous studies, only a small subset of these features were considered for detecting specific types of event and topic. This study presents an algorithm that takes these features into account, extends the scope beyond topics and events to detect news topics.

Other contributions of this research include a thorough review of the literature in microblogs, concerning text models, topic and event detection techniques, and related tools for gathering information from Twitter.

Rich amounts of data in microblogs open up a lot of opportunities. Successfully implementing a news detection system will bring these significant outcomes.

- Successful representing tweet using patterns allows us to discover higher quality and informative topics. These topics can then be mapped into higher-level feature space such as ontology and semantic web, and advance the knowledge discovery process for deriving intelligence.

- Topic detection and news identification from microblogs enable us to find more news topics that might otherwise be missed, which provides a cutting edge to discover emerging news from citizen journalism. These topics can be combined to complement with information from other mainstream media, to potentially enhance reading experience and improve the performance of personalization and recommendation systems.

- A microblog news detection system can be implemented during disasters and emergencies to monitor the event development and provide timely reponse and necessary support. It can also help the public services to increase situational awareness of the

environment during accidents, deploying rescue teams to provide assistance at the earliest possible time.

1.5 Thesis Outline

The structure of the rest of this thesis is illustrated in Figure 1.2.

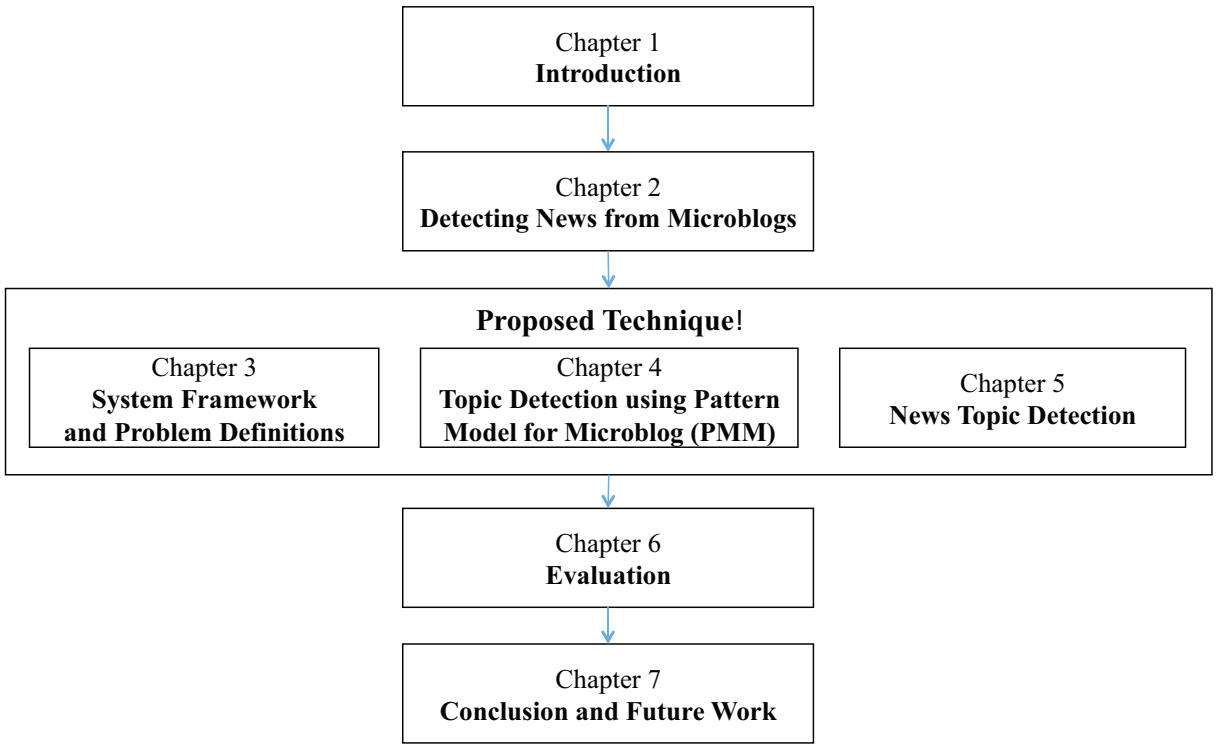


Figure 1.2: Thesis Structure

Chapter 2 reviews literature in the area of detecting news from microblogs. We investigate state-of-the-art systems, tools and applications that are currently used to find news topics from microblogs. We use Twitter as the main study platform; therefore we can also investigate different tweets-related properties and analysis techniques. Chapter 3 presents our research, introducing the system framework and formally defining this research.

Chapter 4 outlines our topic detection technique. We present our Pattern Model for Microblogs (PMM), which utilizes sequential patterns to extract topics from microblogs. We also present algorithm to eliminate noises and redundant patterns. Chapter 5 presents

our news detection algorithm, which evaluates the news value of topics using multiple features.

In Chapter 6, we formally evaluate our models and algorithms. This chapter describes the evaluation methodology, experimentation design, dataset and presents the experimental results with in-depth discussions. Chapter 7 concludes the research project and thesis, and highlights the recommended future work and directions.

1.6 Publications

The refereed publications are part of the key results from the research, work presented in this thesis has been previously published in international conferences and journals.

Lau, C. H., Tao, X., Tjondronegoro, D., and Li, Y. (2012). Retrieving information from microblog using pattern mining and relevance feedback. In *Data and Knowledge Engineering*, volume 7696 of *Lecture Notes in Computer Science*, pages 152–160. Springer Berlin Heidelberg

Lau, C. H. and Tjondronegoro, D. (2010). Text mining in microblogs for real time topic and event monitoring. In *Super Computing (SC'10) Early Adopters PhD workshop*, New Orleans, USA

Lau, C. H., Li, Y., and Tjondronegoro, D. (2011). Microblog retrieval using topical features. In Voorhees, E. M. and Buckland, L. P., editors, *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011,*, Gaithersburg, Maryland. National Institute of Standards and Technology (NIST)

Tao, X., Zhou, X., Lau, C. H., and Li, Y. (2013). Personalised information gathering and recommender systems: techniques and trends. *ICST Transactions on Scalable Information Systems*, 13(1-3)

Tjondronegoro, D., Tao, X., Sasongko, J., and Lau, C. H. (2011). Multi-modal summarization of key events and top players in sports tournament videos. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 471–478

Chapter 2

Detecting News from Microblogs

This chapter reviews related work and techniques in microblogs news detection from the content, characteristic and technique perspectives. Twitter, one of the popular microblogging platforms today, is particularly investigated. We discuss features used in Twitter processing and evaluate current techniques to detect topics and events from Twitter. To conclude the chapter, we present a summary of gaps in the research which will be addressed in this thesis.

2.1 Why Microblogs?

Microblog is a new medium emerging together with the rapid growth of social media and mobile technologies. Java et al. (2007) define a microblog as

> *a form of blogging that lets you write brief text updates about your life on the go and send them to friends and interested observers via text messaging, instant messaging (IM), email or the web.*

Microblogs enable users to participate, exchange and share information using web links, comments and personal opinions. Rich amounts of news-related information are available, but at the same time are accompanied by large amounts of personal babbles (Java et al., 2007).

While microblogs have become popular only in recent years, the history of microblogging can be traced back to 2005, when the term *TumbleLog* was first coined[1]. Microblogs came into greater use later when services such as Tumblr and Twitter arose. Table 2.1 lists the popular microblog service providers. Jaiku, initially launched in 2006 and bought by Google in 2007, was terminated in early 2012 due to the low number of users. Other platforms include Identi.ca and Plurk, each serving different target audience with slightly feature differences. Weibo (direct translation of *"microblog"* in Chinese) attracts a large number of users and attention from academic research (Gao et al., 2012), with Sina[2] and Tencent[3] as the main providers. Among all, Twitter is the most popular microblogging platform, with users across different domains of expertise including celebrities, sports players, national leaders and the general public.

Service	**Key Features**
Twitter [4]	Designed to simulate the idea of using SMS services to communicate among a small group.
Jaiku [5]	Thoughts and comments driven services in regards to users' lives and other subjects
Plurk [6]	Strong focus on conversation, timeline view and emoticons
Identi.ca [7]	Microblog service built on open-source standards and tools
Weibo	China based microblog services mainly targetting Chinese users

Table 2.1: List of microblog service providers and key features

The term *"microblog"* often leads users to think of it as a scaled-down *"micro"* version of blog (Gamon et al., 2009), but they are two fundamentally different media. Blogs are designed primarily for authors to provide details of comments and opinions of their own expertise. Writing blog post is time consuming, as authors need to ensure content quality and validity. On the other hand, microblogs are designed to be short and concise for fast dissemination, and to be compatible with the mobile network. Microblogs allow information to be transmitted in real time, promoting user participation in event

[1]http://www.kottke.org/05/10/tumblelogs
[2]http://www.weibo.com
[3]http://t.qq.com

discussion and live reporting. Microblogs lowers the need for time and cognitive load for content generation, which leads to a higher update frequency. A blogger might be updating a blog once every few days, whereas a microblogger may post several updates in a single day.

Many studies reveal usages and behaviors of the microblog, highlighting its use as an alternative news source. The length of a standard microblog message is approximately the same as that of a typical newspaper headline and subheading, which makes the microblog an ideal medium for news sharing (Naaman et al., 2011).

Much news is broken on Twitter before being reported by public media. One of the significant cases is that of the death of the superstar Michael Jackson in 2009. At 2:26pm on 25th June 2009, the news was reported and spread virally on Twitter but was identified as a hacker attack by Google[8]. Users were unable to verify the news validity until 25 minutes later, when Google finally confirmed that Michael Jackson had died. Only then did the mainstream media start to report the news (Kaplan and Haenlein, 2011).

In 2008, when *Mars Phoenix* (a Mars exploration spacecraft) found evidence of water, NASA chose to release the official announcement through their Twitter account, instead of at a formal press conference. This surprised many major news outlets, which picked up the news only hours later[9]. This shows that Twitter is becoming a platform not only for citizen journalism reports to emerge, but also for major organizations to release important news.

While Twitter is well known for spreading *"yellow news"* and soft press, it can be used in serious matters such as political incidents. Twitter was used wildy during the tenth Iranian presidential election when Iranian authorities restricted access to popular social media websites such as Facebook and Youtube (Burns and Eltham, 2009; Grossman, 2009). Mobile communication channels were blocked to prevent protesters from sharing violence footage to the public. Twitter, which was not blocked, subsequently became an

[8]http://www.dailymail.co.uk/sciencetech/article-1195651/How-Michael-Jacksons-death-shut-Twitter-overwhelmed-Google–killed-Jeff-Goldblum.html

[9]http://mashable.com/2013/09/23/twitter-history-moments/?utm_cid=mash-com-fb-main-link#gallery/twitter-tweets/52402e4797b2f86aaf001dea

important medium for users to share photos of demonstrations and protest. These items of content are shared virally on Twitter and later on picked up by media companies and journalists and publish on the mainstream media.

2.1.1 How Microblogs Affect News Production

The rising popularity of Twitter has not only affected news media with its ability to discover breaking news, it has also influenced the news production cycle. Twitter emerges as the primary channel for journalists to disseminate information during major events such as the 2008 American election (Huberman et al., 2009), the 2009 Iranian Election (Grossman, 2009) and the 2011 Egyptian Revolution (Starbird and Palen, 2012). Journalists use Twitter to engage with audiences, track news development, and promote their work (Ahmad, 2010; Hermida, 2010, 2009). News media also actively use Twitter to post articles as an additional channel for publicizing news report[10].

Twitter is designed for viral distribution by nature, which can be seen from the word limits that facilitate its dissemination using mobile devices (Ienco et al., 2010). Users actively share information, engage breaking news and provide live coverage of events (Farhi, 2009). Microblogs create a new form of journalism that is developing together with the way Internet is influencing media production (Hermida, 2010).

The rise of multimedia elements and digital work has aggressively reshaped the job scope of a journalist in the professional sense. Twitter opens a different landscape by enabling technological assets like hypertext, multimedia, and interactivity (Steensen, 2011); this has affected journalistic norms and practices (Deuze, 2005). This change, pressuring journalists through the extended work cycle has inevitably raised questions (Lievrouw, 2005; Deuze and Marjoribanks, 2009).

Apart from using microblogs in their daily work routine, journalists also use MuckRack.com, a dedicated platform to connect with other journalists and communicate among themselves. Journalists on social networking platforms always provide insights

[10]http://www.journalism.org/analysis_report/news_agenda_twitter_vs_traditional_platforms

and sideline reports of the stories, or how the stories are curated. They also offer personal opinions on news events, which is different from the factual reporting ways traditional news are presented. This provides an additional context for the development of news coverage, and additional transparency with news (Hayes et al., 2007). All these active developments have raised concerns and ethical issues, causing news organization leaders to adopt social media policies to bring social media usage in line with professional journalism practice (Hermida, 2010, 2009).

While traditional media are still the market leaders in news reporting, journalists are facing high pressures in the competition of reporting news at the earliest possible time, while at the same time making decisions on what to publish. Online service and social media hence become the key factor that affects journalists' decisions. Citizen journalism enabled by mobile technology has further contributed to the pressures that journalists are facing, leading journalists to review their workflow and the need for advance systems in their daily operations (Deuze, 2005).

Citizen journalism is considered a bottom-up journalism activity and that has made its way into formal journalism because of three factors: (Bruns and Highfield, 2012):

1. The rise of Internet as a mainstream medium led to substantial increase in websites for specialist purposes.

2. The substantial increase in possible channels of news content has led to journalistic organizations reducing resources in news production, which in turn lowers the average quality of journalistic production.

3. Conservatism in mainstream journalistic operations limit the engagement between journalist and readers.

Microblogs have not only helped the audience to obtain news from multiple perspectives more easily, it has also enabled users to become active in the news creation process in which messages move back and forth and users have a chance to interact with information (Stassen, 2011). Journalists can easily source not only for news, but for

sideline information such as complaints and comments about a product or brand, actively contributed by the community (Jansen et al., 2009).

2.1.2 Specific Characteristics of Twitter

Twitter is one popular microblogging platform that has many unique characteristics. One interesting finding highlights that using Twitter is similar to writing diary back in eighteenth century (Humphreys et al., 2013), where:

1. They are both semi-public in nature.
2. They are both introspectively chronicling activities and trivial day-to-day logs.
3. They both exist in narrative form.
4. The entries are rather short.

The main differences are the lack of social interaction and the inability to systematically subscribe to or "follow" one another.

Twitter allows users to compose messages (called *tweets*) to send to a network of associated *followers* using a variety of devices. A tweet allows only 140 characters, which is approximately the length of a typical newspaper headline and subheading (Milstein et al., 2008), compatible with the existing short messaging system (SMS) of mobile phones. Short tweets are convenient for users to compose, consume, and communicate their thoughts anytime and anywhere.

Today, Twitter has gained popularity, with over 200 million registered users, over 180 million unique visitors daily, and delivering1600 tweets sent per second on average (Yarow, 2010) . Twitter includes users from different fields, consist of celebrities Lady Gaga (*@lady-gaga*) and Justin Bieber (*@justinbieber*), national leaders Barack Obama (*@barackobama*) and Kevin Rudd (*@kevinrudd*), news publishers CNN (*@cnn*) and Associated Press (*@ap*) as well as the general public.

One key question always being discussed is whether Twitter should be characterized as a "social network" or as "news media"? This is because much news-related information is spread within Twitter network, but it also models a certain level of social relationship. Evan Williams (founder of Twitter.com) described Twitter[11]:

> *What we have to do is deliver to people the best and freshest most relevant information possible. We think of Twitter as it's not a social network, but it's an information network. It tells people what they care about as it is happening in the world.*

There are a few users with more than a million followers; they are either celebrities (e.g. Ashton Kutcher, Britney Spears) or mass media (e.g. Ellen Degeneres Show, CNN, New York Times). Celebrities and politicians are users who have more than 10k followers). Users with more followers are found to be likely to tweet more frequently.

2.2 Finding News from Twitter

The large volume of tweets makes finding news topics from Twitter challenging for both humans and machines. We still lack useful tools to help users to locate news-related topics. Existing tools have some applications for keyword monitoring and trends analysis, but are not specifically designed as tools for news.

2.2.1 Review of Existing Tools

Currently available solutions for finding news topics in Twitter are mainly designed for visualizing trends and counting occurrence of keywords, hyperlinks, and hashtags. Several applications have been developed for finding and tracking topics (Table 2.2). These systems process incoming tweets to provide real-time statistics but do not effectively solve the information overload problem. Most of the systems still require human effort to achieve satisfactory result.

[11] http://blogs.cornell.edu/newmediaandsociety2010/2010/02/24/twitter-i-dont-care-about-your-daily-minutiae/

Users who wish to discover news topics exploratively can find out current trending topics from Twitter.com, or can use directories such as *wefollow, twellow* to find popular Twitter users. Such an approach suits users who do not have an event or topic in mind and who wish to only find out the on-going discussion. Users need to manually inspect the tweet content of each trending topic just to find out what are the actual topic behind the tweets. The Twitter homepage provides a list of trending keywords with preliminary filtering to remove stopwords and commonly appeared words. Twitter uses a proprietary algorithm to show 10 keywords or phrases that behave as "trending" characteristics. This sometimes reflects current events ("nba final") but often gives just frequently appearing keywords such as *#ladygaga* or *#tgif*, without differentiating the nature of the trends.

Twitter provides only simple search functionality with keyword matching. Search results are presented in reverse chronological order, assuming users are interested in the latest information. Google and Bing attempt to index and integrate tweets during a real-time search to complement the web search results, but the service was terminated soon after it was launched as they have not found a reliable way to combine web search and microblogs[12].

Users who already have a topic in mind can utilize services such as HootSuite to monitor selected keywords and hashtags. However, this method will only monitor and provide visualization of tweets volume over time, without any in-depth analysis. Manually selected keywords might not be the actual keywords or hashtags used by the other Twitter users who experiments multiple combinations in order to widen the coverage.

Few systems provide statistical insights for Twitter trending topics. Trendistic[13] provides a list of trending topics from Twitter with chart visualization, but the topics are restricted to frequently occurred terms without any filtering. This leads to meaningless terms without much semantics (e.g. *cool, lol, fun*) superceding other important topics. TweetMeme[14] monitors hyperlinks embedded in the tweets and measures popularity based on the number of times it has been mentioned. These systems have been discontinued

[12]http://mashable.com/2011/07/04/google-realtime-search-suspended/
[13]http://trendistic.com
[14]http://www.tweetmeme.com

Service	Service Type	Description
Twitter Trends	Trends Analysis	Provide top 10 trending topics or hashtags from Twitter.com
TweetStats	Trends Analysis	Charting tool for trends by user
Trendistic†	Trends Analysis	Daily chart based on query keywords
Monitter†	Keyword Monitor	Allow users to monitor selected keywords
WeFollow	User Ranking	Directory of prominent user based on prominent scores
Twellow	User Ranking	List of popular users in different category based on number of followers
JustTweetIt	User Ranking	Directory of users using predefined category
TwapperKeeper†	Data Collection	Crawl Twitter API to collect tweets in exportable format
HootSuite	Data Collection	Paid service to crawl Twitter API and collect tweets using keywords and hashtags
Twitter Search	Search	Search and return results in reverse chronological order based on query
Google Real Time Search†	Search	Search and combine results from Twitter within the original webpages search

† = service no longer available.

Table 2.2: Existing tools for finding or tracking information from Twitter

since Twitter no longer allows third party services to collect, process and redistribute tweets.

In summary, human effort is still required to find news information. Multiple attempts are mandatory to experiment with different combination using keywords, hashtags and users to widen the coverage. The efficiency of these tools is still undetermined as most of the tools are third-party solutions and their performance depends on the traffic allowance decided by Twitter[15].

[15]https://twitter.com/tos

2.2.2 News Detection Systems

In general, a news detection system is capable of processing a set of documents and discovering the latent topics underneath. A typical news detection system (Figure 2.1) is a three steps process: (i) extract the distinctive features that represent the content; (ii) detect the set of topics within the input documents; and (iii) identify news-related topics.

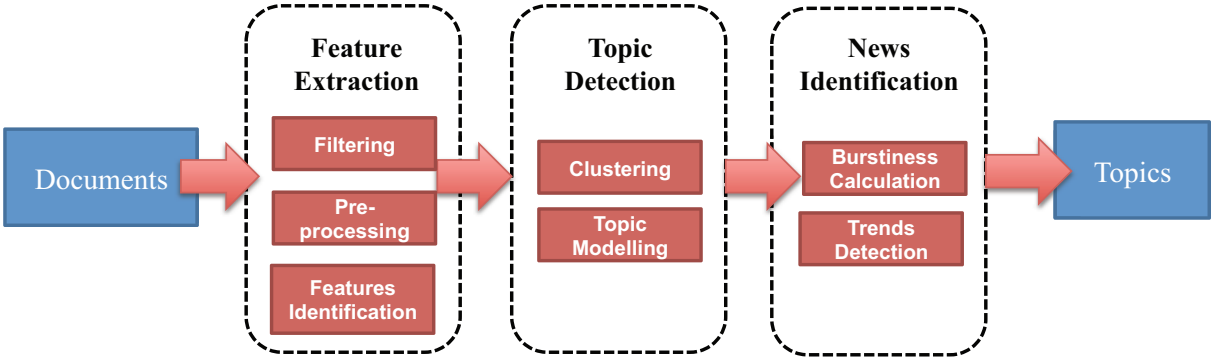


Figure 2.1: Typical news detection framework

A news detection system is different from news aggregators such as Google News and Yahoo News. A news aggregator performs Topic Detection and Tracking (TDT) to process the large amount of information from newswire, radio and television, to find new events and track its development. It then allows users to receive regular updates about the trending news topics.

News detection works by processing large number of documents and aims to find out if there are any interesting topics that are related to real-world events. BlogPulse scans 100,000 weblogs daily to find trending topics (Glance et al., 2004). News detection is difficult to perform on microblogs, given its volume and the amount of noise.

Some notably microblog news detection systems are described below. Eddi is an interactive topic browsing system for Twitter (Bernstein et al., 2010) which groups tweets into topics and allows the user to visualize topics using tag cloud. TwitterStand (Figure 2.2) processes news-related tweets and allows users to browse news according to geographic focus. TwitInfo (Marcus et al., 2011) allows users to explore events visually. As Eddi is only applicable for a single user stream and does not process tweets from the public stream, this research will focus on comparing TwitInfo and Twitterstand.

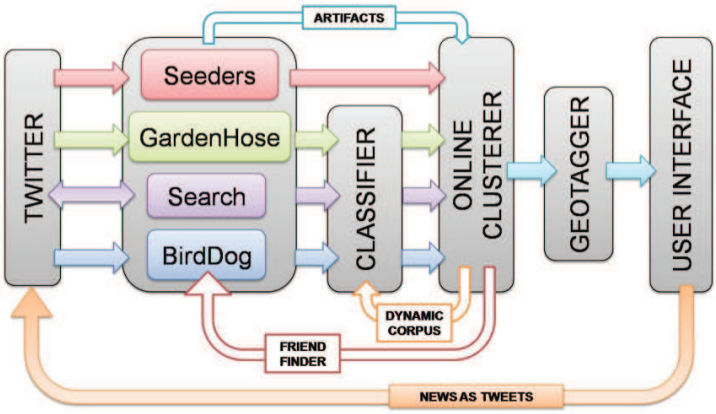


Figure 2.2: System architecture for TwitterStand (Sankaranarayanan et al., 2009)

From the architecture perspective, TwitterStand and TwitInfo have more complete feature sets, including a tweet crawler, an event identifier and a topic identifier. Twitinfo is designed to assist users in browsing events specified by users. While TwitterStand does not require a query, its topic detection performance depends on the pre-selected seeders list.

TwitterStand performs online clustering using term frequencies to find topics and periodically merging duplicate clusters. TwitInfo detects topics by calculating the peaks from time series tweets frequency. Both systems use TF-IDF weightings to reduce the effect of popular terms. Twitinfo further applies sentiment analysis to assist users in visualizing key moments in an event, and TwitterStand uses social relationships to introduce additional seeders.

2.3 Extracting Features from Twitter

Microblog processing is studied extensively by Efron (2011, 2010), Efron and Golovchinsky (2011) and Efron et al. (2012), who point that the short length of text has caused a subtle difference in tweets. This has affected the performance of feature extraction in Twitter, leading to the failure of many existing text models. Extracting distinctive features thus becomes an important task in microblog processing.

Feature extraction in data mining is the primary step for extracting distinctive information that characterizes the data. In text mining, feature extraction breaks down an input document and translates natural language to a machine understandable form. Feature extraction is crucial as it directly affects the performance of subsequent tasks.

Text REtrieval Conference (TREC) dedicates a track to investigate adhoc search and identify useful features from Twitter (Ounis et al., 2011). The short length of tweets has resulted noisy, ungrammatical, full of abbreviations (Table 2.3), and misspellings content. Special Twitter syntaxes that require special processing further increase the difficulty in feature extraction.

Acronym	**Expanded Form**
AFAIK	As Far As I Know
DM	Direct Message
RT	Retweets
IMHO	In My Humble Opinion
PRT	Please Re-tweet
NTS	Note to self
CRE8	Short version of *"create"*

Table 2.3: Common acronyms used in tweets

Content and Context are the two main features that can be extracted from tweets (Figure 2.5). Text content can be processed by text analysis tools to extract significant terms; sentiments can be detected by checking the opinionated words. The Twitter community has also invented different syntaxes for spreading information among users and promoting tweet searchability (Liao et al., 2012). These Twitter-specific features can be used to infer further insights to improve content understanding and assess tweet quality (Naveed et al., 2011b).

Additional information can be extracted from a tweet's metadata for various purposes. Username identifies individual users and can be used to eliminate topics that are dominated by single user, which are prone to be spam. The following and followers relationship can be utilized to verify user popularity, and estimate the information diffusion. User

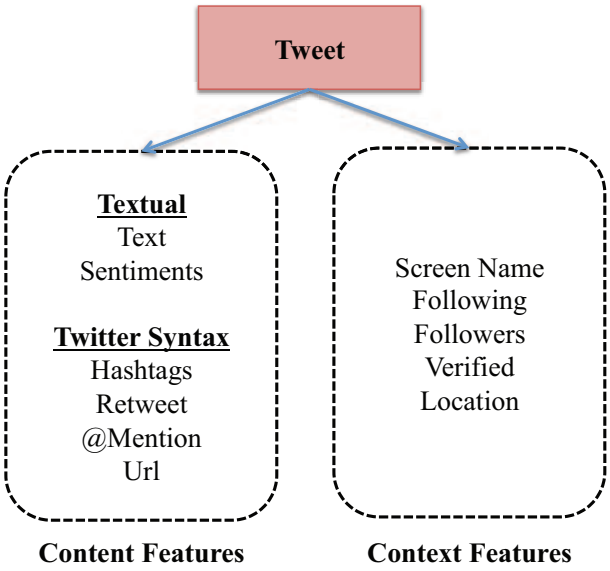


Figure 2.3: Type of features in a tweet

verification can be used to filter and rank tweets (Uysal and Croft, 2011); user location can determine interesting topics of a specific geographic(Sankaranarayanan et al., 2009).

2.3.1 Term Frequency

Term frequency, the most common content feature used in text processing, indicates term importance in a document (Feldman and Sanger, 2006; Salton and Buckley, 1988), as shown in Figure 2.4. It is also the primary feature in many microblog applications (Table 2.8) because of its simplicity and maturity in implementation (Naveed et al., 2011b; Massoudi et al., 2011).

One problem of using only term frequency is that all terms in a document are considered equally important, although this is not always the case. Some terms have less discriminating power for relevance. For instance, football-related articles are likely to contain *"football"* in almost every article, but the terms *"English"* and *"American"* provide more information about the actual type of *"football"* an article belongs to.

To address the problem, Inverse Document Frequency (IDF) is used to measure discriminating power and specificity for terms in document collection. IDF reduces the effect of frequently appearing terms by scaling down weights for terms with high

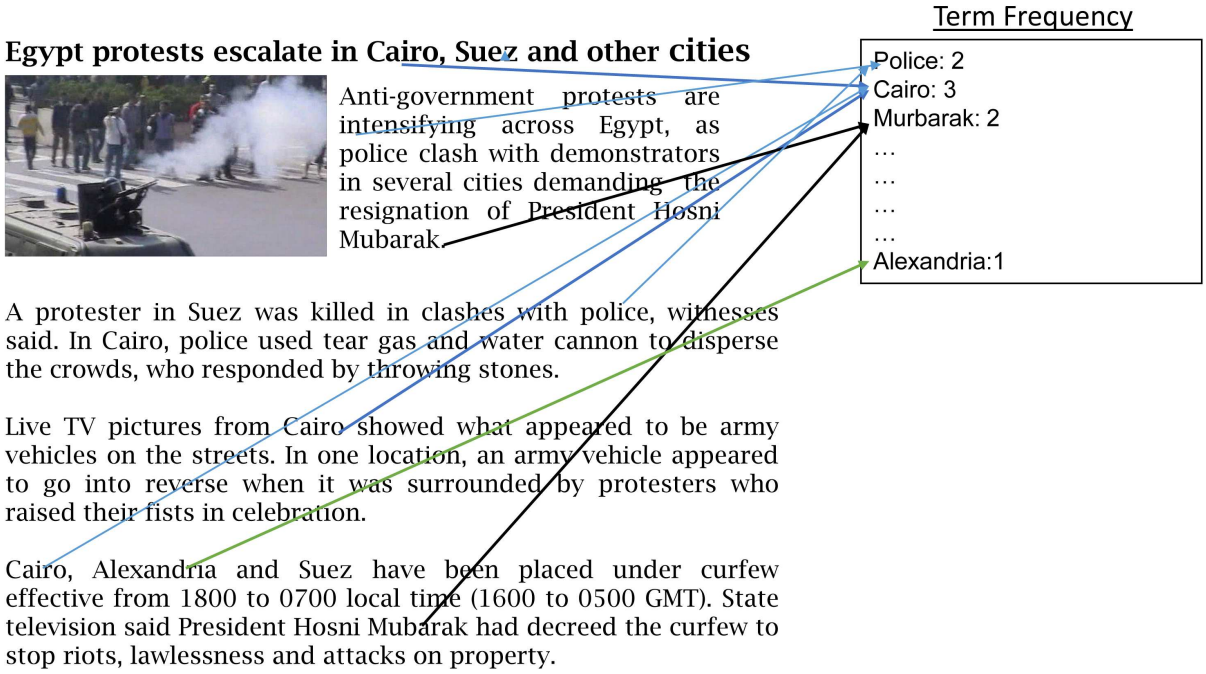


Figure 2.4: Example of Term Frequency Weighting

appearance using the equation below:

$$IDF(t) = log\frac{|D|}{DF(t)}$$

where D is a set documents and document frequency $DF(t)$ represents the number of documents where t appears. IDF is usually combined with TF to indicate term importance in a document, defined as:

$$TFIDF(t) = TF(d,t) \times IDF(t)$$

Existing IR models rely highly on term frequency to indicate importance, but the performances drop when being applied to microblogs (Efron et al., 2012). According to Naveed et al. (2011b), this is due to the low verbosity in tweets, which means if TFIDF is applied directly, the weights are essentially just IDF, since the term frequency are always close to binary (either 0 or 1). Therefore TF-IDF is unable to signify the distinction between tweets. Statistics from the TREC microblog dataset show that about 80% of the tweets contain terms that appear only once; another 10% of tweets contain terms that

appear 2 to 3 times; terms rarely appear more than 3 times in most tweets. Even if a term occurs more frequently, it does not necessarily carry more semantics, as shown in Table 2.4.

ID	Tweet
1	Our Generation Has Had No Great Depression, No Great *War*. Our *War* Is A Spiritual *War*, And Our Depression Is Our Lives
2	Job security is at a premium. *Train* yourself or *train* your replacement? You choose! Take The Dog Test and find out how to *train*...
3	Bachmann To Give Her Own State Of The Union Rebuttal *WHY WHY WHY*!!!
4	*Oprah* Winfrey: What is Her Family Secret?: By Lyneka Little *Oprah* Winfrey's Family Secret: *Oprah* Winfrey
5	Big Cuts At *BBC*: The *BBC* is to re-shape *BBC* Online by 2013 to deliver its public service mission

Table 2.4: Example of tweets containing frequent terms

While IDF can normalize the length effect of a full document, it behaves differently in microblogs. A high *df* can means an important topic word as it appears in many tweets, reducing the importance of such terms may lead to a drop in performance performance (Lee et al., 2011a). Using IDF in tweets is also not reliable as noisy and rare terms will have a higher IDF score. It is also impractical to constantly recalculate IDF as tweet collection grows rapidly.

Another popular term weighting approach in IR is the Okapi BM25 model, defined as

$$score(Q,D) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,D) \cdot (k_1+1)}{f(q_i,D) + k_1 \cdot (1-b) + b \cdot \frac{|D|}{avgdl}}$$

BM25 is a scoring function that ranks a set of document D using query Q. The function considers document length ($avgdl$), and free parameters k_1 and b can be determined by optimization (by default k is within [1.2, 2.0] and $b = 0.75$). BM25 has also been used as an aggregating function to group similar tweets for event detection (Albakour et al., 2013).

Term frequency methods are effective but calculating global weights is still impractical,

as tweets arrive at a furious rate and many terms have a short and bursty life span, arriving and diminishing quickly within a short period. Term weights show only the importance of a term within a collection, not its importance with respect to time. Therefore term weight alone is insufficient for finding time-sensitive topics such as news.

2.3.2 N-gram and Phrase

One key deficiency of term-based models is that the order of terms is not captured, as documents are represented by Bag-of-Words (BOW). BOW representation considers a document as a unordered terms collection, which leads to information loss during processing, making users have difficulty understanding the end result. An n-gram model captures the feature as a *"pattern"* of expressions, preserving neighboring words and sequential information. For instance, a bigram can be a person (*"Barack Obama"*) or a product (*"Apple Ipad"*), A trigram can then be a phrase such as *"search engine optimization"*. The n-gram model provides more semantics to terms *Barack* and *Apple*, which is useful for capturing topics and memes in microblogs.

N-grams can be used to address the common *term mismatch* issue in IR. Term mismatch happens when terms used in documents do not match any of the query terms, although the document is related to the query. Two term mismatch problems are *polysemy* and *synonyms*. Polysemy refers to words that carry multiple meanings such as *"apple"* which can means fruit, or the technology company *Apple Inc.*; the term *"bank"*, which can mean the border of a river, a financial corporation (e.g. *Bank of Queensland*), or a location (e.g. *South Bank*).

Synonyms refer to multiple words that share the same meaning or are used interchangeably by different communities. For instance, *taxi* and *cab* both refer to public transportation, *hat* and *cap* both refer to the head accessory. It is also common for tweets of a particular topic to use different keywords, causing data sparseness and lack of context. Table 2.5 shows relevant tweets returned by query term *'war'*, but of different meaning. Some words are used interchangeably (e.g. *football*, *soccer*) and may carry different meanings. Tweets are composed by different keywords to refer to the same

concept. For example '*riots*' and '*protest*' are both used to describe the citizen disorder at Egypt (Table 2.6). Users are unable to specify their actual query intention, leading to inconsistent retrieval results and causing users to perform multiple iterations to widen search results.

ID	Tweet
1	The God of *War* Internet Games is a hot seller in eBay. Buy it.
2	just reached level 20 on Global *War* Online on my iPhone! Click the link to join my squad #iphone #ipod #ipad #Gl
3	The Call: 2011's Top Risks: In Mexico, the drug *war* heats up: More broadly, the political consen... #tcot #tlot #p2
4	@stuartimmonen Did he buy it from you? I didn't look at the Star *Wars* stuff, I'm not a big *war* in the stars dude.
5	@cd78 GnR - Welcome to the Jungle? Nirvana - SLTS? Black Sabbath - *War* Pigs? Deep Purple - Highway Star? - Cochise is good though!

Table 2.5: '*war*' with different meaning in tweets

ID	Tweet
1	RT @alaa: disturbing news from egypt *riot* police withdrawing frm many posts witnesses saw police spill gasoline on cars downtown #Jan25
2	Thousands in Egypt denouncing Mubarak, clashing with *riot* police; demonstration is biggest in years
3	@Yemen2011 Hope it stay calm. In Egypt the Masjids are being surrounded by *riot* troops and vehicles. Oh man pls no Black Friday.
4	Violent Egypt *protests*: US urges peace as violent *protest* against President Mubarak in Egypt continues.
5	Long Live Anarchy! @OWFreeNation (Winter):RT @OWKL: After Egypt: What happens if this anti-authoritarian people?

Table 2.6: Tweets retrieved using '*riot*' and '*protest*'

Another area that n-gram has shown effectiveness in microblogs is sentiment analysis. Bigrams and trigrams can be combined to improve the sentiment analysis performance of micro-reviews and movie reviews, compared with using unigram only. By combining

the unigram, bigram, trigram and stopwords, the analysis performance is significantly improved (Bermingham and Smeaton, 2010). Tweetmotif shows that bigrams and trigrams support users in performing exploratory search and can be used to summarize topics for thematic and faceted search (O'Connor et al., 2010b).

Essentially the idea behind n-Grams and phrases is to capture more information for overcoming the limitation of terms, and modelling the sequential relationship between terms. Sequences can capture key elements in news such as person, location and products. Text mining techniques such as Frequent Pattern Mining (FPM) can be used for a similar purpose but only limited studies has conducted. However, the initial results are promising (Lau et al., 2011, 2012).

2.3.3 Hashtags, Retweets, Mentions

Along the way, Twitter users have invented different conventions to support tagging, forwarding and communication between users. Three prominent syntaxes that are used and now integrated into Twitter's core feature are hashtags, retweets and mentions.

Hashtags

Hashtags allow users to assign short labels to indicate the topic of a tweet. This practice is similar to the tagging behaviour in other social media (e.g. Delicious, Flickr) that provides textual explanation of media content. Hashtagging is similar to the concept of a chatroom in Internet Relay Chat (IRC), for forming discussions around a particular topic (Chang, 2010). A hashtag, a single term with a '#' sign prefix denoting the category of a tweet, forms discussion from casual topics *'#tgif'* to specific events such as *#emmys, #olympics* and *#eqnz* (Table 2.7) .

Hashtags play an important role for users to share content and to disseminate information. They provide additional context and facilitate tweets search by grouping similar tweets. Without hashtags, a tweet can be seen only by one's own followers, with no other ways to be found easily unless there are matching keywords. This limits the ability

of tweets in spreading information. Hashtags promote searchability and allow a tweet to reach further and to be seen by more users. Using hashtags encourages users to participate in discussion and allows event organizers and authorities to gather information around particular topics and events (Xiao et al., 2012).

Hashtag	Event
#qldflood	Queensland Flood
#ausvotes	Australia Federal Election
#egypt	2011 Egyptian revolution
#iranelection	Iranian election protest
#emmy	Emmy Award
#nbafinals	NBA Basketball Final

Table 2.7: Example of news event related hashtags

Hashtags can also detect interesting topics in microblogs. Weng et al. (2010b) propose a community-based model to discover popular hashtags by measuring users' dispersion and divergence using graph-based theory. Correa and Sureka (2011) recommends tag from tweets to connect to other social networking such as Flickr and Youtube, and Xiao et al. (2012) recommend news topic oriented hashtag using term co-occurrence. These approaches find only interesting hashtags and recommend only the interesting contents, but do not ensure the topics are of users' interest.

Hashtags are also used to improve retrieval performance (Efron, 2010) and sentiment analysis (Davidov et al., 2010). They are used to include more terms in query expansion and to improve retrieval performance (Efron, 2011, 2010). For news-related applications, Abel et al. (2011) use hashtags to model Twitter user profiles and provide personalized news recommendation. Hashtags are heavily used during events including Egypt Protest (Papacharissi and de Fatima Oliveira, 2012), riots in UK (Vis, 2013) and epidemics (Culotta, 2010a). This shows that the hashtag is a good feature for locating news topics from microblogs.

Retweets

Retweet, a user action that forwards a tweet from one user to that user's own followers, is indicated by *"RT"* at the beginning of a tweet. It can be used to engage readers, or to invite users to participate a thread without directly addressing them (Boyd et al., 2010). Retweeting shows a user implicitly expressing their recognition, level of agreement and finding a tweet interesting or important (Kwak et al., 2010).

Retweet activity can be used to discover usage patterns in different situations. During crises and disasters, retweet is heavily used to disseminate important information and media (Bruns et al., 2012). A retweet can be used as an indicator for estimating topic interestingness (Naveed et al., 2011a).

In some cases, users tend to discard retweet information, not attributing the original author, especially where a tweet is about action, or crowdsourcing, for example:

> *Join @MarkUdall @RitterForCO and @BennetForCO to support an up or down vote on the public option http://tr.im/Cm2u.*

Nevertheless, users always credit the original authors when the original tweet contains informative content, such as hyperlinks, videos and images, which helps to understand a topic (Nagarajan et al., 2010).

Mentions

User mention is a syntax to refer specific users or to address one another in using *@username* (or *@mentions*) form (Boyd et al., 2010). Using @mention implies communication or attention seeking, and a two-way mention represents a conversation. The @mention is also known as @replies when used as a form of address to indicate recipients intended to gain target user's attention while posted in public (Honey and Herring, 2009). For instance, a user who writes "*I am watching @oprah now!*" is trying to get Oprah Winfrey's attention and hoping that she will reply.

The measurement of in-out degree property of @mentions can also determine the different roles users are playing within a community (Bruns et al., 2012). A transport department Twitter account may receive many inbound mentions when users are actively

reporting traffic conditions, while a helpdesk operator is tweeting to many different users to answer their enquiries. Mentions can also be used to locate social information, to find individuals with specific interests, or to figure out other users' view on a specific topic (Teevan et al., 2011).

2.3.4 User feature

User level features exist in all social networks to indicate the relationship between users. However, the relationships between Twitter users are relatively weaker compared with other social media such as Facebook, Youtube and Flickr, according to a study byKwak et al. (2010).

In the same study, Kwak et al. (2010) surveyed 41.7 million Twitter user profiles, 1.47 billion relationships and 106 million tweets to study the network properties of Twitter. On the social relationship aspect, most of the celebrities and mass media do not follow their followers back, resulting in a low level of reciprocity (i.e. User A follows User B and User B follows back User A). There are 77.9% of users connected in a one-way direction and only 22.1% have reciprocal relationship. Much higher reciprocity is found on other social media websites: Flickr (68%) and Yahoo 360 (84%). The low reciprocity in Twitter is because Twitter does not require users to follow their followers back in order to communicate, where other social networks such as Facebook and LinkedIn, users are required to be connected before they communicate with each other.

Interestingly, another localized survey by Weng et al. (2010a), based on top-1000 Singaporean Twitter users listed in *Twitterholic.com*, shows the opposite. The study reveals that 72.4% of users follow more than 80% of their followers and 80.5% of the users have 80% of their friends follow them back. Twitter also demonstrates a network property that diverges from traditional social media networks. The distribution of followers is not power-law (this can be because the users do not need to endorse their followers). On average, the users have a short degree of separation and the following/follower relationship does not really model the actual connection between the two users.

User relationship can be utilized as a feedback mechanism to improve data quality. TwitterStand constantly monitors different tweets source, comparing the followers among users (Sankaranarayanan et al., 2009). This is based on the idea that users who follows many other news contributors are likely to be interested in news. A user who follows many news contributors and who is also contributing sufficient news-related tweets will be considered a trustworthy information seeder.

Most studies rely on graph theory, which requires large amount of computation, reducing the scalability when applied to news detection system. User following and followers changed from time to time. There is no solid evidence showing a strong correlation between a user sharing a topic being related to the users they are following.

Similarly, users who have a large number of followers are well-known public figures such as celebrities and politicians. Tweets from these users will generally receive lots of attention and get retweeted very quickly. Although verified users show a certain level of information credibility, this can show only that the information is credible, not necessarily a news topics(Castillo et al., 2011).

2.3.5 Sentiments

Sentiment analysis has long a history. It has been previously studied for different platforms such as blogs, forums and reviews (Pang and Lee, 2008), and has only recently been extended to microblogs. Sentiment analysis extracts attributes and components of a particular subject that have been commented on, and classify the nature of these comments as positive, negative or neutral (Liu, 2008). The subject can be a product, person, event, organization, or topic. It can be associated with components and subcomponents, each with its own set of sentiment attributes.

Sentiment classification can be categorized into supervised and unsupervised methods. Supervised methods use different aspects of text features as their training set. Pang and Lee (2008) classify movie reviews using supervised machine learning techniques including Naive bayesian, Maximum Entropy and the Support Vector Machine (SVM),

using unigrams, bigrams, part-of-speech and term frequencies as features. Kim and Hovy (2006) extract opinions, opinion holders and topics expressed in online news using Natural Language Processing (NLP). Godbole et al. (2007) extend the value of named entities in news and blogs by indicating their level of expressivity.

Subjective opinion in microblogs is an important feature that differentiates microblogs from traditional text documents. As tweets are short and topic-focused, it is sensible to perform sentiment analysis on tweets to understand user's opinion on the associated topic. This task is relatively easier than for other media such as blog and news. Blogs usually contain mixed sentiment opinions from many perspectives, news articles are factual report which should not contain much of the personal emotions of the reporter.

Twitter is a rich corpus for sentiment analysis and opinion mining, Pak and Paroubek (2010) and Nielsen (2011) publish a list of sentiment terms, targetting microblogs only, that extends the Affective Norms for English Words (ANEW) sentiments list. Bermingham and Smeaton (2010) applied a supervised training method to classify sentiments in tweets, and reported that the short length of tweets actually helps in classifying sentiments, compared with full-length blogs. Brody and Diakopoulos (2011) show that terms commonly considered as informal language (e.g. niceeeeeeee, cooooooool, goooooood) bear strong indications of sentiments, and such intentional lengthening can imply importance.

While sentiments cannot be applied directly to detecting news topics, it adds an extra dimension to complement the understanding of public feelings and reaction level towards a topic (Barbosa and Feng, 2010). Topics with controversial sentiment polarity indicate that users with passionate discussion, care about the topic and therefore make comments to express their opinions. Cheong and Lee (2011) study the effect of microblogs in terrorism informatics during the 2009 Jakarta and Mumbai attacks. The study shows that Twitter data can be turned into a decision-making tool by coupling sentiment information with data mining techniques.

For political events, Shamma et al. (2009) show that tweets can be used effectively to annotate and describe events using opinion data, and that such statistics can produce cues on venue, structure and the activity level during presidential debates. Sentiment

analysis has been used to determine the effects of polls (O'Connor et al., 2010a). It has also been shown that tweet sentiments correspond closely to voters' political preferences and correctly reflect the election results (Tumasjan et al., 2011). Public opinions were used to capture topics in the live presidential debate between Barack Obama and John McCain (Diakopoulos and Shamma, 2010).

Public mood and opinions in diverse fields (e.g entertainment, politics and economics) can be extracted using sentiment analysis. Stock market price has been found to directly correlate with tweet sentiments (Bollen et al., 2009) and can be used to predict trends of stock (Bollen et al., 2011). The focused nature of the tweets and the high density of sentiment-bearing terms benefit various automated sentiment analysis techniques in microblogs (Bermingham and Smeaton, 2010).

2.4 Finding Topics from Twitter

The majority of Twitter users aim to find timely information about news topics, trending topics, and events summaries. The intention behind is to understand *"what was happening?"* and why a topic is trending (Teevan et al., 2011). Users are also interested in updates for real-time topics including local incidents, traffic and the status of online services.

Topic detection is the fundamental task in Twitter news topics detection (Figure 2.5). Volume of tweets in each topic will then be evaluated to measure its "burstiness" to identify topics that are trending. Trending topics will then be filtered to select topics that are related to events. News scores for the events are then computed, leaving only topics that are related to actual news events.

Traditionally, Topic Detection and Tracking (TDT) is the main method for finding topics from news articles, using retrospective event detection and online new event detection (Allan et al., 1998). TDT identifies news stories from corpus streams such as news articles and large chunks of text with no separation (e.g. transcribed speech from broadcast news).

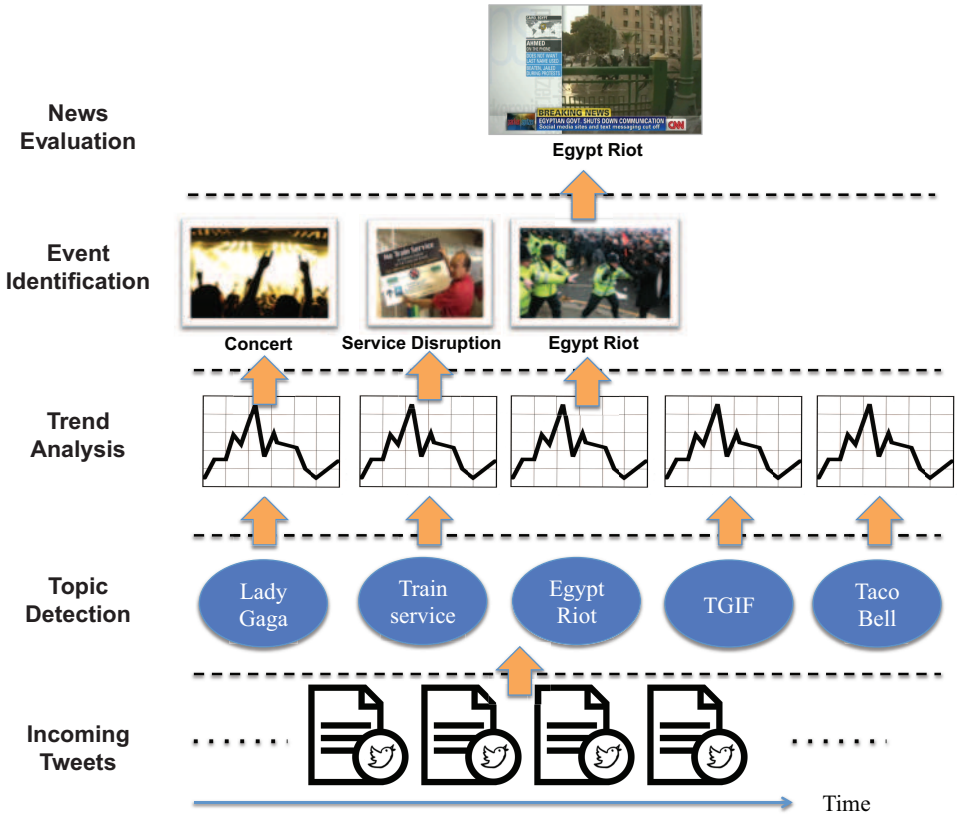


Figure 2.5: News Topic Detection Process in Microblogs

In TDT, event detection is performed restrospectively to group stories from the corpus into clusters. Each cluster represents only one event, and one news article can belong to only one cluster. Topic tracking then associates incoming articles with events that happened earlier. First Story Detection (FSD) then identifies the first instance of a new event that has recently happened.

Substantial amount of work that has been conducted for TDT (Kontostathis et al., 2004). In particular, topic detection on full news articles (Allan et al., 2005, 1998) and social media such as blogs (Sekiguchi et al., 2006) has achieved great success. TDT is effective for news articles, as the incoming source is well-formatted and rich amounts of news information are available. The challenge is to detect the boundary and find the first instance of particular news. TDT techniques are still not readily applicable for microblogs, as microblogs have varying content qualities and structures requiring special treatment and processing techniques for topic detection (Java et al., 2007; Kwak et al., 2010).

2.4.1 Techniques Overview

Emerging trends in Twitter aim to provide users a quick overview of what are the hot topics attracting the attention of a majority of users. These topics are often driven by emerging events and breaking news. The topics are obtained using terms frequency, and users with large follower bases (e.g. Justin Bieber, Lady Gaga) or common hashtags (e.g. *#nowplaying*—denote what songs users are listening now) often dominate top positions in trending topics. This problem has been solved only recently[16]. Trending topics are still limited to 10 topics and do not provide users with much insight for understanding or determining whether the topic is news relevant.

Many studies use the terms *"bursty topics"* and *"trending topics"* to describe the emerging hot topics detected from Twitter. In a broader sense, the characteristics between bursty topics and trending topics are similar. In order for a topic to become trending, it will have to demonstrate certain levels of *"burstiness"*, which shows the sudden increase "spike" in frequency.

2.4.2 Trending Topic Detection

Trend is the basis of a trending topic for finding emerging topic by performing trend analysis. This has been done before on blogs in BlogPulse (Glance et al., 2004); Cheong and Lee (2009) analyze trending topics in Twitter to identify patterns for decision making.

The Twitter homepage provides a list of top 10 trending topics at a specific point of time using proprietary algorithm. The trending topics are keywords with high occurrence but users are unable to define the time period and locations. The topics provided are usually terms and hashtags, which require users to find out more information and details themselves, in order to understand why the topic is trending. While there are third-party services such as WhatTheTrend[17] that utilize crowd sourcing to attempt to provide definition for trend, such information is not always reliable, relying on human effort to provide semantics on a trend, and does not have any significant evidence to show that it

[16]http://mashable.com/2010/05/14/twitter-improves-trending-topic-algorithm-bye-bye-bieber/
[17]http://www.whatthetrend.com

Approach and Domain	Model	Technique Description
(Nichols et al., 2012) Sports	Bursty keyword	Extract multiple sentence summary using word frequency.
(Shamma et al., 2009) Politics	Newton's method Social graph	Use peak detection and network graph to analyse microblogging practice in live events.
(Mathioudakis and Koudas, 2010) Trending topics	Bursty keyword PCA	Detect and merge multiple bursty keywords as trend.
(Sharifi et al., 2010) Trending topics	Phrase	Provide single sentence summary for a given topic.
(Sakaki et al., 2010) Earthquake	Temporal model Spatial model	Use Twitter user as sensor to estimate earthquake location.
(Culotta, 2010b) Epidermics	Linear regression	Detects influenza using multiple regression models.
(Liao et al., 2012) Epidermic, Trends	Epidermic simulation	Identify events by simulating epidermic model, grouping similar users based on links and identify trends in events.
(Goorha and Ungar, 2010) Products	Bursty keyword Contextual statistic	Identify interesting phrases and cluster related terms to form emerging topics of products.
(Quincey and Kostkova, 2010) Epidermics	Terms co-occurrence	Provide indication of possible flu outbreak.
(Paul and Dredze, 2011) Health	Topic Aspect Model	Track illness and measuring risk factors by using prior knowledge.
(Rowe and Stankovic, 2011) Conference	Proximity clustering Naive Bayesian	Align tweets with events and sub-events using semantic annotation.
(Xu et al., 2012) Hot Topics	TF-IDF Apriori	Detect hot topics from Chinese microblogs using data mining techniques.
(Schulz et al., 2013) Accidents	Naive Bayesian JRipper SVM	Detect small scale incidents to improve situational awareness.

Table 2.8: Comparison of topic detection technique, application and feature used in microblog topic detection

correlates with news topics.

Trendsmap[18] provides map-based visualization of trending topics. Users can sign up for alerts for latest trend updates. Similarly, Twendr's dashboard view provides a geographical view for trending topics in each country. Trendistic provided frequency visualization for comparing terms, but the service is discontinued.

Emerging trends on Twitter can be classified into the exogenous trend, which includes broadcast events, global news, important days, physical events, and the endogenous trend, which includes memes, retweets and fan activities (Naaman et al., 2011). The key features for collecting content aggregation statistics for trend analysis are content, interaction, participation, time and social.Wilkinson and Thelwall (2012) compared the top trending topics of English tweets from nine countries (United Kingdom, United States, India, South Africa, New Zealand and Australia) using 0.5 billion tweets. The study shows that festivals and religious events are most popular, followed by media events, politics, human interests and sports. Users are most concern with trending topics from U.S and least concern with topics from India. Such imbalanced findings explain the perceived importance of an international hierarchy, which echoes the news coverage and which might cause the media to become overpowered.

Cataldi et al. (2010) proposed a 5-steps process to model the life cycle of a term using a novel aging theory based on user authority, calculated using PageRank algorithm. The emerging term selection is based on nutrition (term quality) and energy (term burstiness). Twitter trends can also be mined with data mining techniques such as Kohonen's Self-Organising Map (SOM) to visualise user demographic of trending topics, revealing the underlying pattern and characteristics for decision making (Cheong and Lee, 2009).

2.4.3 Bursty Topic Detection

The concept of the *bursty* topic is an important concept to detect unusual *"surges"* in text stream, which is a key problem in temporal text mining (Kleinberg, 2003). Finding

[18]http://www.trendsmap.com/

spikes from the text stream identifies events by tracing terms and phrases in a time series. This concept is extended to detecting *"bursty topics"* from microblogs (Zhao et al., 2010; Du et al., 2011).

The idea of finding bursty patterns from data streams is not new. Kleinberg (2003) propose a state machine to model the arrival times of document in streams and find bursty data. Ihler et al. (2006) model a sequence of count data using Poisson distributions. The limit of these methods is that the data stream must represent only a single topic.

To consider the time factor in term frequency, Lee et al. (2011a) proposed a sliding window model which computes term weights based on its arrival rate within a given time frame. The weighting scheme is able to detect the shift and concept drift by considering temporal factors when deciding term importance. This model is proven effective for modelling relationships between events (Lee et al., 2012). Although the presented keywords can represent temporal characteristic, the topics extracted still require manual analysis for further understanding.

Bursty topics can be used to reveal events that attract public attention. Du et al. (2011) detect bursty topic using term weights together with users, number of followers and replies. Lee et al. (2011a) use a sliding window to find the bursty topics. Burstiness can be used to quantify how trending a term is during a specific timespan of events happening (Metzler et al., 2012).

Weng and Lee (2011) detect bursty topics by characterizing words patterns using a wavelet, and group these words into topics. However, this method does not scale up well when the number of words becomes bigger. While the model from Diao et al. (2012) captures more terms in a topic than that of Weng and Lee (2011), which contains two to three words per topic, the remaining problem in the two models are topics still represented as a set of terms, which is hard to understand without interpretation.

Diao et al. (2012) based their model on the observations that posts around the same time are likely to share the same topic, to detect event related posts. Posts published by the same user are likely to share the same topic and so can help to filter out personal posts. This study shows that unique topics can be detected from bursty topics using

temporal and user interests as the main features.

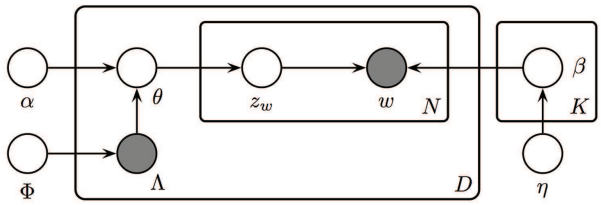
The bursty score for a term during a time period can also be used to generate event storylines (Lin et al., 2012). TwitterMonitor detects bursty terms that arrive at unusually high rate, and groups co-occuring keywords together to form trending topics (Mathioudakis and Koudas, 2010). Bursty terms are identified using queuing theory; Principal Component Analysis (PCA), Singular Value Decomposition (SVD) and Named Entity (NE) Extraction are applied afterwards to derive contextual information and trend description.

2.4.4 Topic Modelling

Topic models have been applied in wide range of areas for identifying latent semantics from text content. Latent Dirichlet allocation (LDA), one of the popular techniques for its performance and flexibility (Blei et al., 2003), is a generative model that presents underlying topics as a set of infinite mixture. Each document is considered as a probability distribution of topics; its probabilities can then be estimated through sampling methods.

One variation of LDA, Labelled LDA (L-LDA), is used to model tweets for ranking and user recommendation task (Ramage et al., 2010). L-LDA mixes labelled topics together with latent topics discovered using original LDA Figure 2.6. While performance of TF-IDF in microblogs retrieval is questionable, Ramage et al. (2010) show that TF-IDF and L-LDA performed similarly in ranking task, and that TF-IDF outperforms L-LDA in the user recommendation task by about 30%. A combination of L-LDA and TF-IDF improves the ranking by a slight 3% but significantly boosts the performance in the user recommendation task by 66%.

Conversely, various studies suggested that LDA does not work well on Twitter due to the short length of tweets (Hong and Davison, 2010; Weng et al., 2010a). One idea to overcome such a problem is to group tweets together to provide more context. Tweets can be grouped by content using terms (Hong and Davison, 2010), by topics (Weng et al., 2010a) or by users using the author-topic (AT) model (Rosen-Zvi et al., 2004).



For each topic k in 1..K, draw a multinomial distribution β_k from symmetric Dirichlet prior η.
For each tweet d in 1..D:

1. Build a label set Λ_d describing the tweet from the deterministic prior Φ
2. Select a multinomial distribution θ_d over the labels Λ_d from symmetric Dirichlet prior α.
3. For each word position i 1..N in tweet d
   a. Draw a label $z_{d,i}$ from label multinomial θ_d
   b. Draw a word $w_{d,i}$ from word multinomial β_z

Figure 2.6: Graphical illustration of Labeled LDA and description of the generative process (Ramage et al., 2010)

However, studies show that direct application of the AT model does not yield significant improvement, compared to the simple term-based approach (Hong and Davison, 2010; Zhao et al., 2011).

In the Twitter-LDA model, Zhao et al. (2011) show an important finding—*"a single tweet is usually about a single topic"*. The authors show that content aggregation performs better than author-based aggregation. This might be due to less variation in content-based aggregation than in author-based, and signals that Twitter authors do not necessarily tweet about the same content.

2.5 Identifying Event and News from Twitter

The terms *"events"* and *"topics"* are sometimes used interchangeably in microblog research, as most of the real-world events are topics, but all topics are not necessarily referring to real-world events. This section describes only the studies that are monitoring real-world events.

Twevent segments tweets into semantic phrases to facilitate human interpretation, while bursty segments are identified within a fixed window period based on frequency

patterns (Li et al., 2012). The newsworthiness of a segment is then computed by its correlation with Wikipedia. Weng and Lee (2011) use wavelet transformation and auto correlation to measure the burstiness of each word. Terms with high energy are retained as events. Event similarities are measured by cross correlation between event segments, but this approach does not scale well.

The Locality Sensitive Hashing (LSH) technique can be used to extract local news events from Twitter (Agarwal et al., 2012). The long tail of local news events can be stored as *event-object* by filtering buzzwords for different events; For example, factory fire may contain *"fire", "blaze", "factory", "plant", "mill"*. Prior knowledge is applied in supervised classification and boosting to discard irrelevant messages. Each potential event-object can then be merged using time-of-occurrence and location.

Historical events can be discovered retrospectively using structured representations (Metzler et al., 2012). Text mining and semantic web can be combined to detect accidents from tweets (Schulz et al., 2013). Long et al. (2011) show that topical words can be clustered together to represent events; For instance, an event containing *"ipad2"* and *"apple"* suggest its relationship with the release of Ipad 2. Topical words are selected using document frequency for terms appearing in a document within a day. This approach is similar to the idea of using frequent closed patterns to represent topics.

2.5.1 Critical Situations and Political Events

Critical situations and political events are two significant events at a country level. During critical situations, authorities might shut down communication channels and limit media and journalist access. Social media then become the important tool for information dissemination. For instance, during the Mumbai terrorist attack, Twitter and Flickr were used together as the primary services for eyewitnesses to publish media related to the incident (Lee et al., 2011c). It was shown in the Iranian election in 2009 that it is particularly useful to monitor different dimensions of Twitter data, from a simple message count and top key words to more complex information such as sign up and self-reporting location Gaffney (2010).

For political events such as elections, although not as severe, the fluctuations of public interest regarding different politicians are good indicators for monitoring public response, in order to adjust their election campaign strategy accordingly. Voters can follow the election stories to understand opinions and comments from multiple perspectives. This is hardly achievable by using traditional media. Tumasjan et al. (2011) studied 100,000 tweets and found that Twitter can be effectively used as a platform for political deliverables. Although the discussion is active, about 4% of users contributed to 40% of the total tweets. O'Connor et al. (2010a) show that simple sentiment analysis is adequate to correlate tweets to election polls.

While it has been commonly perceived that Twitter can be used as an effective tool to predict election results, findings from Gayo-Avello et al. (2011) show different outcomes. By using the same dataset (2010 US Senate special election) used by Tumasjan et al. (2011) and O'Connor et al. (2010a), Gayo-Avello et al. found no significant correlation between social media data and electoral predictions, so it might be by chance that Twitter is able to predict electoral outcomes correctly. Authors also noted that the complexity of professional polling cannot be duplicated by simply sampling social media data. The two main factors here are because (i) votes data cannot be collected reliably from social media, and (ii) social media data can be easily manipulated by spammers and propagandists. While there is no solid evidence to show that Twitter is effective in predicting election outcomes, due to the difference in replicating the experiments, there is sufficient evidence from social science study to show that Twitter is an effective tool for gathering public opinion and is powerful enough to cause social change (Bruns and Burgess, 2011; Burns and Eltham, 2009; Gaffney, 2010).

2.5.2 Business Applications

Rich information by users expressing their opinions and comments about products make Twitter a great data source for analyzing consumer behaviours. Jansen et al. (2009) analyzed over 150,000 product-related tweets from consumer electronics, food, services, goods and transportation, on brands including Microsoft, Apple, Starbucks, Dell, Sony

and Adidas. The study shows that automated analysis from microblogs makes no significant difference, compared with manual coding. The linguistic structure of tweets approximates the linguistic structure of natural language expressions, which suggests that microblog analysis can be used as a powerful marketing tool.

Bulearca and Bulearca (2010) further studied the viability of using Twitter as a marketing tool for Small Medium-sized Enterprise (SMEs). They analyze five companies each with two employees from the marketing and public relations (PR) departments. The pilot findings show that Twitter is a critical platform to embark on, and is crucial for the company to gather consumers' opinions directly.

Bollen et al. (2009) consider tweets as *temporally-authentic microscopic instantiations of mood state* and detect six dimensions of mood (tension, depression, anger, vigor, fatigue, confusion) using the extended Profile of Mood States (POMS). The study found that events in social, political, cultural, and economic sphere are highly correlated and have specific effect on various mood dimensions. The technique is extended by Bollen et al. (2011) to measure the correlation between tweets and Dow Jones Industrial Average (DJIA) during presidential election and ThanksGiving day. Goorha and Ungar (2010) process tweets that contain mention of an item (e.g ipod) to uncover subjects deeply hidden in a large amount of information and to extract information that are interested to decision makers, such as brand managers.

2.5.3 Natural Disaster and Epidermic

Analyzing social media during natural disasters has shown that it is able to provide large scale communication involving self-organising behaviour. It produces accurate results, sometimes faster than the official channels (Palen et al., 2009). One key study is the earthquake detection system by Sakaki et al. (2010), which leverages tweets to detect earthquakes and send alerts to communities.

Sakaki et al. consider each Twitter user in Japan as a mobile sensor, computing the probability of an earthquake being tweeted, using time and geolocation information.

Tweets were collected using keywords *"earthquake"* and *"shaking"*, and the posting time and tweet volume were modelled as a exponential distribution to estimate earthquake locations using Kalman filter and Particle filters. It shows that an earthquake can be detected much earlier than the official announcement, even ahead of the impending shockwaves. This provides residents with extra time to take necessary actions and reduce damages.

Qu et al. (2011) studied the effect of an earthquake at Yushu, China in 2010 using Weibo, the main microblogging service in China. The study shows that microblog is useful when dealing with disaster response and the emotion factors can affect the spreadability of message. Lee et al. (2011b) also extends the bursty features from their previous work Lee et al. (2011a), and able to detect earthquake.

Twitter has been used to detect influenza epidemic outbreaks, improving traditional detection in both speed and cost reduction. It complements an existing method that uses Google search and queries data (Ginsberg et al., 2009). Tweets provide more descriptive information compared to search queries; data such as location, gender and age can be derived to provide more demographic insights. Culotta (2010b) detects influenza and Quincey and Kostkova (2010) detected swine flu using pre-defined keywords. These methods detect anomalous change and identify rapid increase in message traffic related to given keywords. The benefits of such a method is to expect more focused information to be collected while listening to the Twitter stream, but the problem is that a set of keywords need to be pre-defined.

A study of the Oklahoma Grassfires in April 2009 and Red River Floods between March and April 2009, found that tweets contained useful and relevant information that enhances situational awareness (Vieweg et al., 2010). The study identifies 10 major features during an emergency, with each feature correlating to different degrees of awareness for different types of event. Bruns et al. (2012) present a qualitative study to identify user behaviour during the severe flood in Queensland, Australia in 2011. Their study shows that hashtags are important during critical events and natural disasters, and that users are actively retweeting information which can be considered as signal amplifiers.

2.5.4 News Event Detection

Twitter has shown many times that it is not just a social media or social network, but a medium akin to traditional news media (Kwak et al., 2010; Lu et al., 2012). However, the amount of tweets generated is tedious for human to process the news information. Therefore it is important to detect and track news topics automatically. There are less attention paid for news detection in microblogs, compared with other media. As seen in previous sections, there are many techniques for detecting emerging topics and events, but not for detecting news events.

Discovering breaking news has always posed a challenge for news organizations. The 24-hour news cycles and real-time social media further complicates the problem of news originating from affected areas to general public. While the credibility of tweets is questionable (Castillo et al., 2011), some Twitter users who encounter early rumours about an unfolding news event will actually search for further information, and share their findings with the community in return (Bruns and Highfield, 2012). Some include appropriate hashtags for the event, which helps the other users who are also interested the event, to become active and sharing information in return.

Previously discussed techniques focused on topics and events, but users are unable to tell which events or topics are related to news. Topic detection covers topics, but does not provide any distinction between news and non-news topics. Topic from a tweet collection can only identify a set of tweets with similar theme. News topics are groups of tweets which demonstrate a unique temporal nature, indicating that an event is happening at a certain time and location.

In general, news topics demonstrate a temporal pattern which they do not exist before a given time, and diminish or become less popular after a certain duration. This research focus only on news with the potential of becoming a news topic, to be published. For instance, the death of Michael Jackson could trigger a spike initially and become an emerging topic, but the subsequent discussions will eventually flatten out. Only the first occurrence of Michael Jackson's death is considered to be news.

One pioneer news detection system for microblogs is TwitterStand (Sankaranarayanan et al., 2009). TwitterStand aims to automatically obtain breaking news from tweets, built upon the idea where *"geographically proximate users often tweet about the same breaking news"*, and therefore limits it effectiveness, making it more suitable for an application with geographic characteristics. TwitterStand relies on an initial set of seeding users who are likely to publish news such as newspapers and television. A naïve Bayesian classifier is then trained to classify tweets between news and junk. Terms in tweets are weighted using tf-idf, and topics are extracted using an online clustering algorithm. Topics with an average tweet publishing time over three days will be discarded. A modified cosine similarity is used to favor tweets with publication times closer to the cluster centroid. Topics are evaluated periodically to prevent fragmentation and duplicates (i.e. multiple topics around the same news, lowering the importance of the individual topic.)

TwitterStand works well in terms of collecting news from microblogs but it is not a completely automatic system. It requires a pre-selection of seed users who are already publishing news content such as newspapers and television. The system also highlights the importance of addressing the problem of important phrases (e.g. *"Barack Obama"*), but in their model preference is given to terms that occurred more frequently (i.e. "Obama") and so it dropped "Barack') as it appeared at a relatively lower frequency.

Other work includes microblogs summarization (Inouye and Kalita, 2011), tweets clustering to aid understanding, focusing on the area of making tweets content easier to understand using manually selected event keywords (Lin et al., 2012). Magdy (2013) presents a news portal by retrieving relevant tweets using pre-defined queries and classifying tweets using SVM to report on popular tweets, jokes, videos, images and news articles. Still missing is a news topics detection technique for microblogs, which would work across different domains, and a single computational model which considers microblog's properties, and not only the bursty terms.

2.6 Literature Evaluation

While microblogs might not replace traditional media and become authoritative sources, they are unquestionably an important alternative with rich information. Microblog services such as Twitter provide real-time information covering multiple aspects of events, and allowing news topics to be discovered.

Existing news detection techniques are built on successful topic and event detection. These techniques detect emerging trending topics that are of user interests, using trends analysis to find *"bursty"* topic; however, they do not indicate which topics are relevant to news events. Topic models can be used to process retrospective data in offline settings, but they do not scale up well in the microblog environment, due to its high computational cost.

Most systems use term-based models that do not consider term relationship, which leads to inconsistent performance. Topics detected using term-based models are difficult to understand, so it can be challenging for users to find information from the results.

Twitter contains specific features that provide unique representation of its content. Hashtags, retweets and urls have all shown useful for topic detection; sentiment information can also be used as a public interest indicator. These features have been applied to dedicated areas only, but not being combined into a single computational model for detecting news topics.

This research tackles the above mentioned issues from two aspects. We apply pattern-based techniques to address problems in term-based techniques when processing microblogs. Then we evaluate the news value of extracted topics, by taking into account Twitter features as well as temporal information and sentiment polarity. The next chapter will present the system architecture which provides high level coverage of components, and formally define the problem for this research.

Chapter 3

System Framework and Problem Definition

Chapter 2 identified the limitations of term-based models and the gaps in existing topic detection approaches. This chapter presents a high-level overview of a microblog news detection framework that aims to address these limitations and gaps. This chapter also formally defines the research problems and notations used in this research.

3.1 System Framework

News topic detection in microblogs is different from traditional topic and trends detection, and from news aggregation. Topics and trends detection identifies emerging topics from tweets, but does not measure the relevance between topics and real-world events. News aggregators such as Google News and Yahoo News focus on news articles, which already contain rich amounts of news topics and a small amount of noise.

This thesis present a news detection framework for microblogs with three main characteristics. The framework is fully automatic and unsupervised, which do not require keyword selection and constant data source updates. In this framework, tweets are represented in pattern feature space to improve performance of both topic detection and news identification tasks. News detection algorithm in this framework considers multiple features, and is not limited to any type of events and domains. Although the framework is designed and tested using Twitter, it can be easily adjusted and mapped to the features

of other microblogging platforms.

The key components in our framework are shown in Figure 3.1. Tweets are collected in Javascript Object Notation (JSON) format using Twitter API and is processed by the feature extraction module to obtain content features and context features. Content features are term tokens from tweets; context features are sentiment scores and Twitter metadata including hashtags, Url, @mentions and retweets (RT).

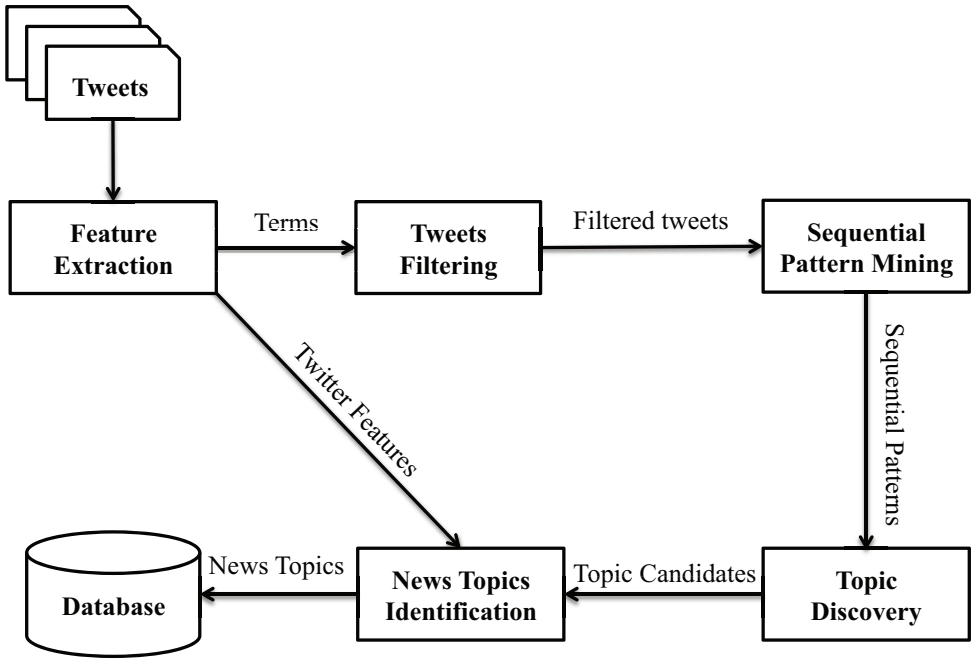


Figure 3.1: Microblog News Topic Detection Framework

Sequential Pattern Mining (SPM) is then applied to identify sequences. The Topic Discovery module then processes the sequences to identify potential news topics. The News Topics Identification module then computes the news relevance score of the topic candidates, and then stores the identified news topics in a database.

3.2 System Framework and Overview

The research focuses on the problem of identifying news topics from tweets. A News topic is defined as "*some incidents that happened at a specific time*" (Yang et al., 1999), and is related to real-world events reported by news media.

Algortihm 1 describes the processing algorithm of our framework. The goal is to detect news topics $\mathcal{N}$, given a set of microblog messages (in this case, *tweets*) $\mathcal{M}$. Tweets are collected during a continous time period from t_i to t_j of an epoch interval (e.g. hourly, daily).

Algorithm 1: Main Algorithm for News Topic Detection Framework

Input: Set of Microblogs $\mathcal{M} = \{m_1, m_2, \cdots, m_n\}$
Output: News topic set $\mathcal{N}$

$\mathcal{P} \leftarrow \varnothing$ `// To store all patterns generated from` $\mathcal{M}$
foreach *microblog* m *in* $\mathcal{M}$ **do**
    `/* Feature Extraction */`
    Process m to extract termset $d = \{t_1, t_2, \cdots, t_n\}$, timestamp v, author username a, hashtags set H, url u, retweet r ;
    Calculate Sentiment Score e, Logit Score l ;
    if m *is noise* **then**
        Remove m from $\mathcal{M}$;
        continue
    end
    `/* Sequential Pattern Mining (SPM) */`
    Generate pattern candidate set P_d for patterns of length from 1 to $length(d)$;
    $\mathcal{P} \leftarrow P_d$
end
Remove non-frequent patterns; `// Pattern pruning`
Closed pattern mining ;
Merge similar topics and construct topic candidate set C;
Calculate news score for $c \in C$;
Construct final news topic set for $score(c)$ above threshold ;
return $\mathcal{N}$

3.2.1 Feature Extraction

We first extract features from $\mathcal{M}$. Term tokens are obtained using standard text-preprocessing techniques to obtain d containg a set of terms t, where $d = \{t_1, t_2, \ldots t_n\}$. Each d is in English language and contains a maximum of 140 characters.

We then extract contextual features of a tweet from its metadata, which contains the following properties:

- Author (a) : Username of the tweet author
- Timestamp (v) : Time when the tweet is posted
- Hashtags (h) : Short label to indicate topic of d
- Urls (u) : URL embedded in d that links to external resource
- Retweet (r) : To indicate whether the tweet is forwarded from another user
- Sentiment (e) : The sentiment score of a tweet

Our model considers only the username; other information such as location, verified user, following and followers count is not considered. Because this other information varies from time to time, modelling such a relationship might not necessarily improve the topic detection performance.

Sentiment has been shown to be an effective indicator in various microblog applications (Sehgal and Song, 2007; Asur and Huberman, 2010), as sentiment analysis benefits from microblog brevity nature (Bermingham and Smeaton, 2010). In the case of news detection, the ratio of positive and negative tweets can be used to indicate public interest level.

The task here is to evaluate the sentiment value e for a tweet d. The sentiment score is also used later in the news score calculation to represent topic polarity by measuring the difference between positive and negative sentiments in a topic.

A tweet message m can then be defined as a tuple:

$$m =< a, v, d, h, r, u, e >$$

3.2.2 Tweets Filtering

Tweets contain large amounts of noise; Therefore, filtering becomes an important step to eliminate irrelevant content earlier and improve information processing performance

(Sharifi et al., 2010). The aim of filtering is to eliminate tweets that are clearly not related to news, such as personal babbles, chit-chats and spam.

We apply a two-step model for the filtering task. Step 1 removes irrelevant tweets based on a set of heuristic rules derived from empirical observations. In Step 2, machine learning is applied to remove noisy tweets using classification model. Removing noisy m from $\mathcal{M}$ ensures a higher quality input for topic discovery.

3.2.3 Topic Discovery

Following the definition in Topic Detection and Tracking (TDT), discovering news topics means finding topics related to people, locations, organizations, events, or phrases, all of which are key elements in news (Allan et al., 1998).

Next we obtain topic candidate set C. We first generate a set of possible patterns P_d from d using Sequential Pattern Mining (SPM). We calculate *support* for every pattern $p \in P_d$, which represents its occurrence frequency in the dataset. We then remove p if $support(p)$ is less than a minimum required value (min_sup). This is to remove noisy patterns that do not carry much representative meaning. The remaining patterns are called *frequent patterns*.

While frequent patterns contain patterns that often appear, there are redundant patterns that contain identical information. For instance, if $p_i =< t_1, t_2, t_3 >$ contains sub-pattern $p_j =< t_1, t_2 >$ and $support(p_i) = support(p_j)$, p_j is redundant and needs to be pruned as the exact information is covered by p_i. Long patterns are preferred as they contain more information with higher specificity. There is no information loss if we remove p_j as we can always derive the information from p_i if necessary.

In some cases, patterns can belong to the same topic but there are slight differences between their supports, which is caused by the noise in the tweets. We will evaluate the coverset of tweets containing these patterns. These patterns will be merged if their

coversets are similar.

The pruned pattern set is now a topic candidate set C, and the next step is to evaluate the news relevance score of each topic candidate.

3.2.4 News Topics Evaluation

Traditionally, topic importance is evaluated using support in a pattern-based model or using term weights in term-based model. The problem of using support is that short patterns are usually general topics that tends to have high support, where long patterns are specific topics but have low support (Wu et al., 2004; Algarni et al., 2009). In term-based models, terms are weighed using Inverse Document Frequency (IDF) to reduce the effect of high occurrence terms. This is based on the assumption that high occurrence terms have less discriminative power since they appear in most documents. However, this is not always the case in news topics detection, as topics that appear frequently are usually hot topics.

Using a pattern model in microblogs means that support can represent oly the popularity of a topic. However, there are four other main factors that can affect topic importance: *burstiness*—representing how trendy a topic is; *hashtags*—representing how active users are tagging the tweets; *retweets*—representing the user agreement level of tweet content; *sentiments polarity*—represents the opinionated level within topics. All these factors are taken into account for evaluating news relevance of the topics. Topics with score greater than the preset threshold will be selected as the final news topics set.

3.3 Chapter Summary

This chapter presents the high level design of our framework and algorithm, which aims to overcome existing gaps of topic detection and term-based models in microblogs, as

identified in Chapter 2. Our framework uses the sequential pattern as the topic model, which captures semantics and models the relationship between terms. Sequential Pattern Mining (SPM) is able to extract key elements in news such as named entities and phrases, and therefore it is a suitable technique for used in microblog news topic detection. However, additional steps need to be implemented to overcome redundant and noisy patterns.

The next chapter describes in detail Pattern Microblog Model (PMM) and related tasks; Chapter 5 presents the computation details for estimating news relevance; performance evaluation of our framework follows in Chapter 6.

Chapter 4

Topic Detection using Pattern Model for Microblog (PMM)

Pattern models have been shown to be useful in many text mining tasks including topic detection, but few studies have been conducted for microblogs. This chapter presents *PMM*, a pattern-based model designed for microblogs. This chapter describes the implementation details and algorithms to overcome issues encountered during implementation.

4.1 Feature Extraction

In data mining, features are distinctive information that represents data characteristics. Feature extraction leverages on the document to derive explicit structure from its implicitly structured representation, but the crucial challenge when applied to microblogs is to extract significant features in a dynamic and noisy environment (Lee et al., 2011a).

From a content perspective, microblogs have a huge term space. Terms emerge and disappear rapidly. Unlike full-length documents, a tweet contains only few words, causing difficulties for statistical information to be gathered. Twitter users often coin new abbreviations or acronyms that are never used in news reports. These terms are not

formal words and do not contribute much towards news topic understanding.

4.1.1 Pre-processing Tweet

Pre-processing removes noisy information that causes negative effects and reduces performance. In microblogs, pre-processing is even more important for their high level of noise. Removing noise can also essentially reduce the feature dimensionality.

The pre-processing is in the order illustrated in Figure 4.1. Part-of-speech (POS) tagging is performed using a POS tagger specially designed for social media, from Carnegie Mellon University (Gimpel et al., 2011). The tagger is capable of identifying special expressions such as emoticons and memes from tweets. POS tagging identifies different parts of speech such as noun, pronoun, adverb, adjective, all of which help to identify useful news topics.

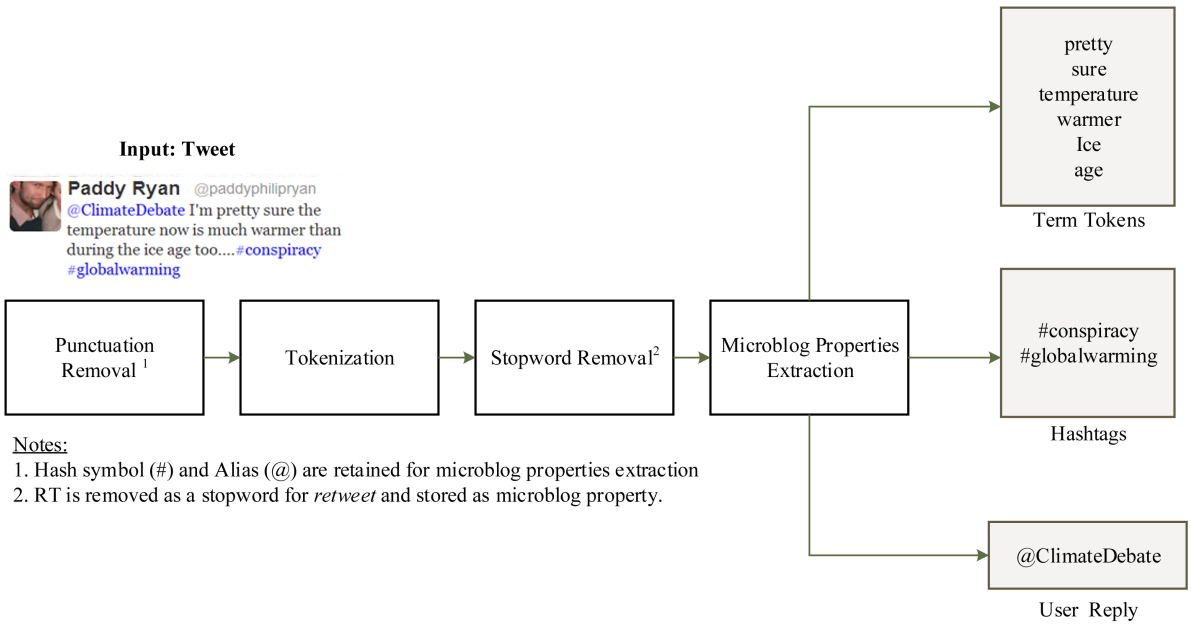


Figure 4.1: Pre-processing Steps

Noise removal and punctuation removal then eliminate non-text symbols and punctuation marks (except @ and #). Tokenization then splits text into individual term tokens. Stop word removal discards words that do not carry indicative meaning and that will

not alter the meaning of a tweet if removed. The stopword list contains common English words such as *he, she, it, is, am, are, the, there*, and a list of bad words (offensive words, swear words) from (Popescu and Pennacchiotti, 2010) and the Google List of Badwords[1]. Lastly, Porter stemming is applied to standardize terms appearance for reducing data sparseness (Table 4.1).

Original Terms	**Stemmed Form**
play, player, plays	plai
conditions, conditional	condition
relations, related, relates	relate
read, reading, reads	read
gent, gently, gents	gent

Table 4.1: Examples of Porter stemming

4.1.2 Filtering Noisy Tweets

Tweets filtering improves topic detection by distinguishing news-related tweets from noise. The goal here is to classify tweets into *relevant* or *irrelevant* with respect to news. Irrelevant tweets are considered less likely to be news related. As it is difficult to differentiate between relevant and irrelevant tweets, our aim is only to remove as many tweets as possible, particularly those that are clearly news irrelevant.

The filtering is in two steps using heuristics and machine learning. We first apply the following rules based on observations:

- Remove terms with unusual length (e.g. *haushaushushahaha, phenylalaninoogl*).
- Remove terms with more than 2 consecutive characters repetitions (e.g. *staaaaaarving, workiiiiing, loveeeeeeeee*).

[1]https://gist.github.com/jamiew/1112488

– Remove tweets with less than three terms.

We then train a classification model using *Support Vector Machine* (SVM) to separate news related tweets and noisy tweets (Han, 2005). SVM is a discriminative model that has proved to be an effective technique for microblog prediction and filtering tasks (Gupta et al., 2012; Tzeng et al., 2012).

The training set for the model contains 1000 tweets, with 80% (800 tweets) labeled as irrelevant, and 20% (200 tweets) labeled as relevant. Relevance label were annotated by human experts from National Institute of Standards and Technology (NIST) manually (Voorhees and Buckland, 2011), and the tweets for training set are selected randomly. This setting simulates the noisy environment in Twitter, in which only a small portion of tweets are related to news.

Table 4.2 shows the features used to build the SVM; Table 4.3 shows examples of the classification results. Tweets contain too many hashtags and mentions are usually considered as noise (d_6). Tweets with short length after pre-processing (d_8, d_9) are obviously irrelevant. Tweets with rich content, together with hashtags or url, are more likely to be related to news (d_1, d_2)

Feature (f_i)	**Description**	**Value Range**
Hashtags	Number of hashtags in tweet	Ordinal
Retweet	Indicate if a tweet is retweeted	Binary
Length	Number of terms in tweet	Ordinal
Url	The presence of URL	Binary
Mentions	Number of mentions in tweet	Ordinal

Table 4.2: Features used for training noisy tweet classifier

ID	Tweets
d_1	Taco Bell sued over beef claims - The lawsuit, filed in California, accuses the fast food chain of false advertising... http://ow.ly/1b1rrv
d_2	Mexican drug smugglers catapult weed over border fence into US: A remote video surveillance system captured dr... http://bit.ly/e9gT8E
d_3	Al Jazeera Breaking: Reuters: Egyptian protestors set fire to govt building & ruling NDP party office in Suez in North Eastern Cairo #Jan25
d_4	Daily Tech: Facebook and Twitter Blocked in Egypt: Both Twitter and Facebook blocked in Egypt during protests -... http://bit.ly/f4mV0N
d_5	BBC News Egypt's violent protests escalate: Egyptian opposition leader Mohamed ElBaradei arrives in Cairo callin... http://bbc.in/hX5Jyh

(a) News relevant tweets

ID	Tweets
d_6	RT @autiizh: RT @vithsz: @autiizh @Dr_Prez @CamsMax @LaughhLikeCyrus @mousion @UKNOWHOITIZ_NYC @RayRally @jaz1421 @Nino_Rob @GRHOP1
d_7	RT RT @Omarvelous Momma said there will be days like this. : there'll be days like this my momma said
d_8	RT @adindacaesarany: There's something new and hopefully success #Februarywish amin
d_9	@Cylister Did Great Today, productive meeting @DrSetItOff
d_{10}	Please follow my friend @Almumuahaha and @Mery_Highland

(b) Non-news relevant Tweets

Table 4.3: Examples of relevant and irrelevant news relevant

4.2 Pattern Model for Microblogs (PMM)

Many topic models in microblog tasks use term-based models, where terms are the main features and topics are presented using the Bag-of-words (BOW) model (Phelan et al., 2009; Cataldi et al., 2010; Zhang and Sun, 2012). BOW does not capture terms relationships, therefore is unable to capture topics such as names, companies and phrases. This has made the topics less interpretable, suffering from term ambiguity and term synonym (Algarni et al., 2009).

Term-based topic detection algorithms project terms into a Vector Space Model (VSM) and cluster documents using standard data mining methods like K-Means. A distance function is then applied to determine similarity between document pairs. This can be hard to determine in large dataset as it is computationally expensive to constantly re-calculate large numbers of documents.

Topic models such as Latent Dirichlet Allocation (LDA) extends Probabilistic Latent Semantic Analysis (PLSA) using a generative model. The main idea behind LDA is that a document is considered as a mixture of topics with different word distributions. Such an idea does not suit short document like tweets, since one tweet usually belongs to one topic (Zhao et al., 2010). LDA is also not easily scalable due to its high computing cost (Zhang and Sun, 2012). Both LDA and K-means requires the number of topics to be defined, which is difficult to estimate in Twitter.

Previous IR studies suggest that in many tasks, pattern-based models achieved better results compared with term-based approaches (Wu et al., 2004; Zhong et al., 2012; Li et al., 2010). Pattern-based techniques such as Frequent Pattern Mining (FPM) discover topics using a set of words to capture co-occurrence and to model the latent relationship between the terms (e.g. *{fifa, world, cup}*). *Support* is then calculated to represent pattern occurrence frequency, which indicates pattern importance, and which can be used to filter irrelevant patterns.

4.2.1 Sequential Pattern Mining

Frequent Pattern Mining (FPM) captures recurring relationships and enables the discovery of interesting correlations between terms. However, frequent patterns do not preserve term order and allow gaps in between terms, which is still ineffective for representing topics and concepts (Li et al., 2010). FPM has also been found to perform inconsistently in a noisy environment like Twitter (Lau et al., 2011).

Conversely, sequential patterns maintain term order and capture underlying semantics, therefore are able to distinguish named entities, companies and phrases, providing more interpretable topics compared with terms (Kim et al., 2012).

Here we present the Pattern Model for Microblog (PMM) using sequential patterns. The implementation details are illustrated in the sample database (Table 4.4). The Sequential Pattern Mining (SPM) algorithm is a recursive algorithm that aims to find the complete set of sequential patterns that covers the topics in a tweet, in their longest form.

Tweet	**Terms**
d_{11}	{*egypt, presid, elect, hosni, mubarak*}
d_{12}	{*egypt, presid, hosni, mubarak, elect*}
d_{13}	{*internet, egypt*}
d_{14}	{*egypt, presid*}
d_{15}	{*internet, egypt, twitter*}

Table 4.4: Sample tweets database related to President Hosni Mubarak and Internet service outage during the 2011 Egyptian election

Given tweet d, contains a set of terms, where $d = \{t_1, t_2, \cdots, t_n\}$. A sequential pattern $p =< t_1, t_2, \cdots, t_n > (t_i \in d, p \subset d)$ is an ordered list of terms generated from d, with duplications allowed. Sequential pattern set P_d for a tweet d can then be generated using Algorithm 2.

Algorithm 2: GenerateSP(d)

Input: tweet d

Output: A set of sequential patterns P_d

```
/* Generate complete set of pattern candidates of all length       */
```

```
1  begin
2      P_d ← ∅
3      for i ← 1 to length(d) do
4          P_d ← Generate(i, d)
5      end
6      return P_d
7  end
8  Procedure Generate(n, terms)
       /* Generate set of patterns of length n for terms              */
9      patterns ← ∅
10     for j ← 1 to length(terms) do
11         pat ← terms[j]
12         for k ← 1 to n do
13             pat ← pat ∪ terms[j + k]
14         end
15         patterns ← patterns ∪ pat
16     end
17     return patterns
```

Table 4.5 shows the topic candidates generated using d_{11} from the sample database. Two patterns of length n and $n+1$ can exist in a super- and sub- relationship, with the following definitions:

Definition 1 (*super-sequence & sub-sequence*)**.** A sequence $\alpha =< a_1, a_2, \ldots, a_n >$ is a *sub-sequence* of sequence $\beta =< b_1, b_2, \ldots, b_m >$, denoted by $\alpha \sqsubseteq \beta$ where there exist integers $1 \leq i_i < i_2 < \cdots < i_n \leq m$, such that $a_1 = b_{i1}$, $a_2 = b_{i2}$, $\ldots a_n = b_{in}$, β is then

Length (n)	**Pattern Candidates**
1	*<egypt>, <presid>, <hosni>, <mubarak>, <elect>*
2	*<egypt, presid>, <presid, elect>, <elect, hosni>, <hosni, mubarak>*
3	*<egypt, presid, elect>, <presid, elect, hosni>, <elect, hosni, mubarak>*
4	*<egypt, presid, elect, hosni>, <presid, elect, hosni, mubarak>*
5	*<egypt, presid, elect, hosni, mubarak>*

Table 4.5: Pattern candidates P_d for d_{11}

called the super-sequence of α.

For example, sequence $< t_1, t_3 >$ is a sub-sequence of sequence $< t_1, t_2, t_3 >$, but sequence $< t_2, t_1 >$ needs not to be a sub-sequence of $< t_1, t_2, t_3 >$ since the terms order is considered. Sequence α is a sub-sequence of β if $\alpha \sqsubseteq \beta$ but $\alpha \neq \beta$.

For a tweet d, we assign *Absolute Support*, $supp_a(p) = 1$ of each pattern p in P_d, as a term mostly appears once in a tweet (Lee et al., 2011a). We define the set of tweets that contain pattern p using *coverset*:

$$coverset(p) = \{d \mid p \in P_d, d \in D\} \tag{4.1}$$

The *Total Support* of a pattern p, is then the number of documents d where p occurs, which is measured as follow:

Definition 2 (*Total Support).* P_d represents the set of patterns of d, where d belongs to tweets collection $\mathcal{D}$. The *Total Support* of a pattern p represents the number of tweets where p appears, denoted as:

$$supp_T(p) = |coverset(p)| \tag{4.2}$$

Table 4.6 shows the calculation of support, represented using n-pattern notion:

Definition 3 (*n-Pattern).* A pattern p of length n is a n-pattern, where $n = length(p)$,

indicating the number of terms contained in p.

For instance, p_a = <*hosni, mubarak*> contains two terms, i.e. $len(p_a) = 2$, and p_a is a 2–*pattern*. A 1-pattern is a special case of $n - pattern$, which is essentially a term.

As we are dealing with large numbers of patterns generated automatically, most of the patterns will not be meaningful and need to be removed. We want to retain only the patterns that frequently occurred, with a minimum number of occurrences. The minimum occurrence can be defined using *min_sup*, and frequent sequential pattern can be formally defined as:

Definition 4 (*Frequent Sequential Pattern)*. A sequential pattern p is called *frequent sequential pattern* if its total support is greater than or equal to a pre-defined minimum support (min_sup in short), i.e. $supp_T(p) \geq min\_sup$.

Minimum support are normally set to 0.2 for optimal performance as per described in other related pattern-based text mining studies (Algarni et al., 2009; Li et al., 2010).

4.2.2 Pattern Pruning

While patterns improve discriminative power, pattern mining process still generates many noisy patterns, increasing the complexity during processing. All pattern-based techniques inevitably generate non-meaningful patterns, either with low occurrence or containing redundant information. These patterns need to be pruned in order to improve performance (Wu et al., 2004).

Pattern pruning removes redundant patterns to reduce the dimensionality of pattern space and decrease the negative effects of noisy patterns. For instance, in Table 4.6, <*hosni*> and <*mubarak*> contain exact information that can be replaced using the single pattern <*hosni, mubarak*>. Minimum support can removes only low frequency noisy

Pattern	Support	Frequent
<*egypt*>	5	Yes
<*presid*>	3	Yes
<*hosni*>	2	Yes
<*mubarak*>	2	Yes
<*elect*>	2	Yes
<*internet*>	2	Yes
<*twitter*>	1	No

(a) 1-pattern

Pattern	Support	Frequent
<*egypt, presid*>	3	Yes
<*presid, hosni*>	2	Yes
<*hosni, mubarak*>	2	Yes
<*internet, egypt*>	2	Yes
<*mubarak, elect*>	1	No
<*presid, elect*>	1	No
<*egypt, twitter*>	1	No

(b) 2-patterns

Pattern	Support	Frequent
<*egypt, presid, hosni*>	2	Yes
<*presid, hosni, mubarak*>	2	Yes
<*hosni, mubarak, elect*>	1	No
<*egypt, presid, elect*>	1	No
<*internet, egypt, twitter*>	1	No

(c) 3-patterns

Pattern	Support	Frequent
<*egypt, presid, hosni, mubarak*>	2	Yes
<*presid, hosni, mubarak, elect*>	1	No

(d) 4-patterns

Pattern	Support	Frequent
<*egypt, presid, hosni, mubarak, elect*>	1	No

(e) 5-patterns

Table 4.6: Pattern and supports derived from sample tweets database

pattern, but does not remove frequent patterns with duplicate information, therefore requires a proper pruning scheme.

Methods such as *maximal patterns* and *closed patterns* have been proposed to reduce redundant patterns. A Maximal pattern is used to obtain the largest coverset of frequent items, but information of each frequent pattern is not retained. We adopt closed pattern as the pruning technique. Closed patterns prune redundant patterns by removing sub-patterns of same support, while still allowing the support of sub-pattern to be derived (if necessary). Using *Apriori* property of frequent patterns, a *sequential closed pattern* ensures maximum coverage by obtaining the longest pattern possible, keeping word order to retain topic semantics. A *closed sequential pattern* is defined in Definition 5.

Definition 5 (*Closed Sequential Pattern)*. A frequent sequential pattern p is a *closed sequential pattern*, if there exist no frequent sequential pattern p', where $p \sqsubset p'$ and $supp_a(p) = supp_a(p')$, $\sqsubset$ denotes the strict subsequence relation of $\sqsubseteq$.

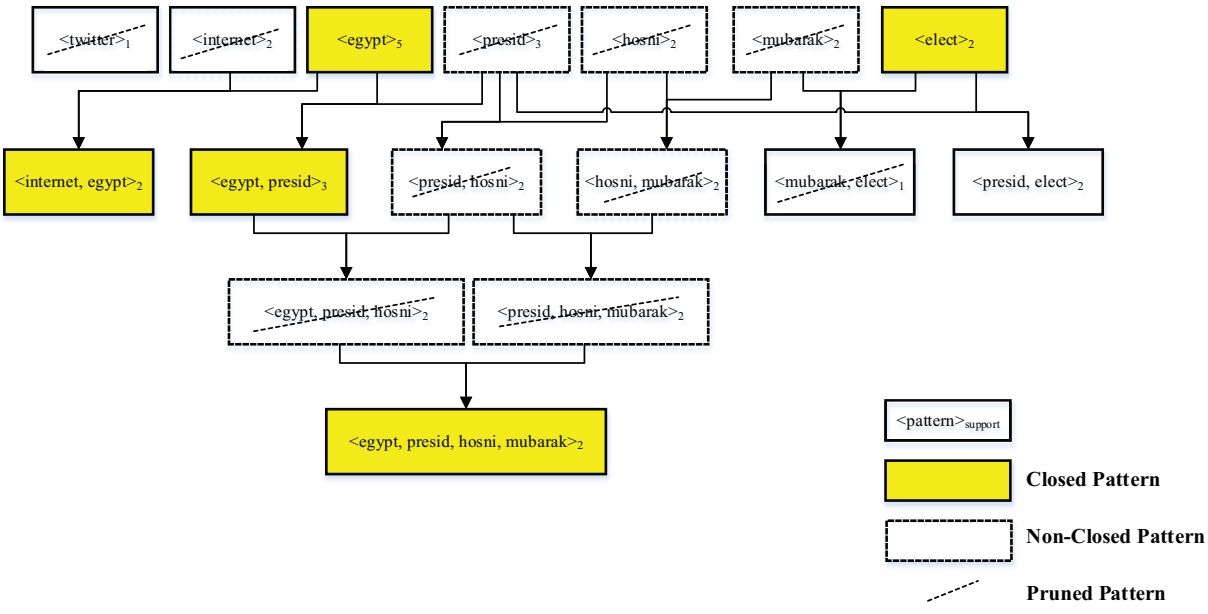


Figure 4.2: An illustration of closed pattern mining process

Figure 4.2 illustrates the process of pruning patterns and we obtain the results as shown in Table 4.7. We can see that the original pattern set containing 16 patterns is reduced to only 5 patterns (31% of the original size). Not only are the topics extracted more descriptive, as for example, *<internet, egypt>*, which describes the internet outage at Egypt. The topics are more readable, as in long topic *<egypt, presid, hosni, mubarak>*, and hierarchical relationship between super- and sub- topics is also maintained.

Once pruning is performed, we obtain a set of closed pattern, the topic candidate set. The remaining problem is that some sub-patterns are not completely merged with their super-patterns. For instance, p_1 is sub-pattern of p_3 and p_5; p_4 is sub-pattern of p_5. These sub-patterns need to be verified and evaluated for further merging.

Pattern Length	**Patterns**
1-pattern	p_1 =*<egypt>*, p_2 =*<elect>*
2-patterns	p_3=*<internet, egypt>*, p_4=*<egypt, presid>*
3-patterns	- nil -
4-patterns	p_5=*<egypt, presid, hosni, mubarak>*

Table 4.7: Pattern set after pruning

4.2.3 Topic Merging

One observation in PMM is that some sub-patterns are not removed during pattern pruning as none of its super-patterns have the same support. For instance, p_4 and p_5 contain similar information and may be further combined. Pattern p_4 is considered as a closed pattern since there is no super-pattern of p_4 that has the same support. Although p_4 is a sub-pattern of p_5, $support(p_5)$ is lower than $support(p_4)$. This is often caused by the noise in tweets where the same topic is described with a slight difference. This not only increases the size of the pattern candidates set, but also reduces the importance of the effect of long pattern p_5.

Merging is required to prevent these topics from propagating and affecting the performance of subsequent tasks. However, there are super-patterns where their support is similar to its sub-patterns. This is usually caused by noise in tweets. The differences between the tweets are trivial, caused by extra terms or slightly different word order. As the sample database shows, d_1 and d_2 belongs to same topic, but position of *"elect"* has caused p_4 and p_5 to remain as two closed patterns, which should be merged.

Topic merging is important for addressing such issue by combining semantically related topics, to provide a better indication of topic importance and to further reduce feature space. In order to determine whether p_i and its sub-pattern p_j are similar, we compute the similarity score between set of tweets which contain the pattern. Similarity between two patterns can be computed as an *overlap score* using Jaccard similarity:

$$overlap(p_i, p_j) = \frac{|coverset(p_i) \cap coverset(p_j)|}{|coverset(p_i) \cup coverset(p_j)|} \tag{4.3}$$

Algorithm 3 describes the pattern merging criteria. For example, to decide whether p_4 =<*egypt, presid*> can be merged with p_5 =<*egypt, presid, hosni, mubarak*>, we evaluate the criteria, using $\delta = 5$ and $\omega = 0.6$:

- Supports of patterns are similar. **Yes**. $\because \Delta = support(p_4) - support(p_5) = 3 - 2 = 1$
- Coversets of patterns are similar. **Yes**. $\because overlap(p_4, p_5) = 0.6$

$$\begin{aligned} coverset(p_4) &= \{d_1, d_2\} \\ coverset(p_5) &= \{d_1, d_2, d_3\} \\ overlap(p_4, p_5) &= \frac{|coverset(p_4) \cap coverset(p_5)|}{|coverset(p_4) \cup coverset(p_5)|} \\ &= \frac{|\{d_1, d_2\}|}{|\{d_1, d_2, d_3\}|} \\ &= \frac{2}{3} \end{aligned}$$

The final pattern space then becomes p_1 =<*egypt*>, p_2 = <*elect*>, p_3 = <*internet,*

Algorithm 3: Pattern Merging

Input: p_i, p_j, support threshold δ, overlap threshold ω; $p_j \sqsubset p_i$

```
1 begin
2     Δ = support(p_j) − support(p_i) ;
3     if Δ > δ then return;
4     O = overlap(p_i, p_j) ;
5     if O ≥ ω then
6         Remove p_j
7         Set support(p_i) = p_j
8     end
9 end
```

egypt> and p_5=<*egypt, presid, hosni, mubarak*>; p_4 is now removed and $support(p_5) = support(p_4) = 3$.

For $< egypt >$ and $< egypt, presid >$, although the support difference is less than 5, since the *overlap* score does not exceed 0.6, they will not be merged.

Lastly, we apply all the patterns to the tweets. In this stage, we want to represent tweets using only the longest patterns that are able to cover all the information without any repetition. This is to prevent common terms within topics (e.g. $< egypt >$) from being calculated multiple times, affecting the topic detection performance. We want to keep only the *maximum pattern* in each tweet, defined as:

Definition 6 (*maximum pattern*)**.** A pattern p_i is called a *maximum pattern*, if there is no other pattern p_j where $p_i \subset p_j$.

The final pattern sets are then applied to the tweets, in pattern space as shown in Table 4.8.

Tweet	Patterns
d_1	<*egypt, presid, hosni, mubarak*>, <*elect*>
d_2	<*egypt, presid, hosni, mubarak*>, <*elect*>
d_3	<*internet, egypt*>
d_4	<*egypt, presid*>
d_5	<*internet, egypt*>

Table 4.8: Tweets in pattern representation

4.3 Chapter Summary

This chapter introduces PMM, a Pattern Model for Microblogs, and describes how to implement PMM to discover topics. We present a novel methodology that extends application of pattern mining to the microblog domain.

First, we describe a process to filter tweets that are irrelevant to news, and extract the content and Twitter features from tweets. To improve topic detection performance, a classification model is trained to separate the tweets that are irrelevant to news.

We then utilize sequential pattern mining to extract topical features. Topics are captured in a natural way. Their co-occurrence and term order relationship are modelled to capture meaningful topics. Pattern pruning is then performed to eliminate noisy and redundant topics, followed by topic merging to further compress similar topics.

The next chapter will describe how to determine the news value of the extracted topics, by calculating a news relevance score using pattern characteristics, context features and other Twitter properties.

Chapter 5

News Topic Detection

The next step after obtaining topics from PMM is to evaluate their news relevance score. Absolute support is insufficient to measure news relevance as short pattern topics tend to overpower the effect of long pattern ones. To address this issue, we first evaluate topic weights using pattern properties and assign topic weights based on their importance in tweets. Then we further calculate the following metrics for each topic: *burstiness* to capture the trending characteristic; *sentiment* to indicate public interest level; *Twitter properties* to model users activities while sharing news tweet. A machine learning model is then built to learn the weighting coefficients of these features.

5.1 Overview

Most of the previous work in microblog events and topics detection used various content (Marcus et al., 2011) and context features (Sankaranarayanan et al., 2009), particularly by the use of burstiness to find trending topics (Lee et al., 2011a; Marcus et al., 2011; Sankaranarayanan et al., 2009). However, there is a lack of formal evaluation that considers the effect of these factors when being applied in news detection.

This chapter presents a news relevance scoring model developed by considering multiple factors (Algorithm 4). We first compute pattern weights to represent content importance, based on its indicative power in a tweet. Pattern weight is then combined with contextual features (burstiness, sentiments) and Twitter properties (hashtags, urls, retweets).

Algorithm 4: News Relevance Score Calculation

Input: A set of patterns P, a set of tweets D_p in pattern feature, weighing parameters α, β

Output: A set of patterns and news relevance score, $\triangle$

$\triangle \leftarrow \varnothing$
foreach *tweet* d_p *in* D_p **do**
    Calculate pattern weight $weight(p)$ for $p \in d_p$;
    Distribute pattern weight to term $t \in coverset(d_p)$;
end
foreach *pattern* p *in* P **do**
    Calculate total weight $W(p)$;
    Calculate burstiness score $B(p)$;
    Calculate sentiment score $S(p)$;
    Calculate hashtag score $HT(p)$;
    Calculate url score $UR(p)$;
    Calculate retweet score $RT(p)$;
    Calculate news relevance score $News(p) = \sum \beta x$
    /* β is the feature coefficients, x is the feature score */
end

5.2 Topic Weight Evaluation

Every pattern is considered as a topic in our model, as most tweets are related to single topic. Therefore, pattern weight organically represent a topic's importance in a document, and the sum of pattern weights across the corpus represents the importance of the topic in a tweet collection. Exisiting pattern-based models measure importance of patterns using support, which is the frequency in which a pattern appears. However, in text mining, longer patterns tend to have low support while short pattern topics often have much higher support (Wu et al., 2004). Such uneven distributions is caused by not taking the pattern specificity into considerations while evaluating support.

There is no doubt that a longer pattern is considered more specific. For example, comparing the pattern *Moscow Airport Bombing* with *Airport* 5.1, the former is more helpful than the latter as a topic, since it carries more information by containing more terms. However, calculating support only implies that *Airport* is equally as important as with *Airport Bombing* and *Moscow Airport Bombing*. The fact is *Airport* is only partially contributing to the understanding of the topic, and the weights need to be properly distributed to reduce the effect of short patterns and model the distinctive power of long patterns.

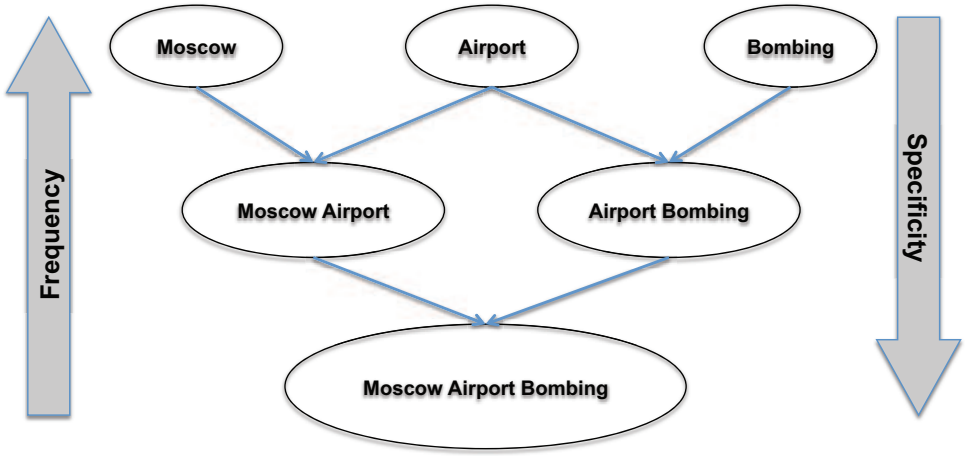


Figure 5.1: Relationship of specificity and frequency between different pattern levels

5.2.1 Evaluating Pattern Weight

The two primary aims of our pattern evaluation method are: (i) to reduce the effect of short patterns, and (ii) assign appropriate weights to long patterns according to the information they contains. We use a tweets dataset to further illustrate the problem (Table 5.1).

The sample dataset contains ten tweets: three tweets (d_1, d_2, d_3) are related to news topic *egypt protest*, two tweets (d_9, d_{10}) are related to *oil price* and *obama government*, and others tweets are irrelevant to news. Most tweets contain a popular term *time*, which is a common phenomenon in tweets processing, with a generic term appears in many tweets.

Tweet	**Patterns**
d_1	{<*egypt*>, <*protest*>, <*fear*>, <*respon*>}
d_2	{<*egypt*>, <*protest*>, <*night*>, <*time*>, <*presid*>}
d_3	{<*egypt*>, <*protest*>, <*presid*>}
d_4	{<*home*>, <*run*>, <*time*>}
d_5	{<*extra*>, <*time*>, <*nba*>}
d_6	{<*time*>, <*year*>, <*again*>}
d_7	{<*time*>, <*year*>, <*valentin*>}
d_8	{<*extra*>, <*time*>, <*parti*>}
d_9	{<*real*>, <*time*>, <*oil*>, <*pric*>, <*obama*>}
d_{10}	{<*obama*>, <*press*>, <*releas*>, <*real*>, <*time*>, <*oil*>, <*pric*>}

Table 5.1: Sample tweets in term feature space

We then perform topic detection using the PMM described in chapter 4 and project the tweets into their pattern feature space using p_{max} notation (Table 5.2). The p_{max} notation represents a tweet using maximum pattern coverage to reduce redundancy when calculating pattern weight in a short document. For instance, d_9 is represented using { <*real, time, oil, pric*>} but not {<*real, time, oil, pric*>, <*time*>} even though both <*real, time, oil, pric*> and <*time*> are frequent patterns. This is different from the traditional representation used in other pattern mining methods (Wu et al., 2004).

Tweet	Patterns
d_1	<*egypt, protest*>
d_2	<*egypt, protest*>,<*time*>, <*presid*>
d_3	<*egypt, protest*>, <*presid*>
d_4	<*time*>
d_5	<*time*>
d_6	<*time*>
d_7	<*time*>
d_8	<*time*>
d_9	<*real, time, oil, pric*>, <*obama*>
d_{10}	<*real, time, oil, pric*>, <*obama*>

Table 5.2: Example of tweets in pattern feature space using p_{max} representation

The problem with using absolute support is that on a document level, there is no statistical difference between long patterns and short patterns. Inspecting d_9, we see that supports for the 1-pattern term <*obama*> and support for the 4-pattern <*real, time, oil, pric*> are both equal to one. Similarly, support for all patterns in d_2, d_3, d_{10} are equal to one. We are unable to distinguish the importance and specificity of patterns in a document.

This problem continues to propagate to dataset level. If we calculate the weight of a topic in a tweets dataset by aggregating the absolute support, we obtain the results shown in Table 5.3.

Topic	Total Support
<*time*>	8
<*egypt, protest*>	3
<*real, time, oil, pric*>	2
<*obama*>	2
<*presid*>	2

Table 5.3: Examples of topic weight calculated using total support

The results show that the popular term *time* overpowers other terms and its support is much higher. This is because generic terms are always used in many tweets. Although long patterns are more specific, in general, they appear less frequently compared with short patterns, hence weakening the importance of long patterns.

To address this problem, we will need to consider the weight of a pattern based on the information it contains. Our idea is motivated by the Pattern Deployment Method (PDM) by Wu et al. (2006). PDM shows that longer patterns are more specific and the support of long patterns can be utilized to evaluate support of term $(1 - pattern)$. Term weights are evaluated based on the term's appearance in long patterns. This method intuitively emphasizes the appearance of patterns, and normalize the noisy effect of general terms that do not carry much indicative power.

The amount of information contained by a pattern can be represented by the number of terms in the pattern. To extract all terms in a document d, we define a *termset* function:

$$termset(d) = \{t | t \in p, p \in P_d\} \tag{5.1}$$

where P_d represents the set of patterns in d. The weight of each term $t \in termset(d)$ is then assigned as

$$w(t) = \frac{1}{|termset(d)|} \tag{5.2}$$

We can represent each tweet $d \in D$ using a term vector representation using (t_m, w_m) notation:

$$\vec{d} = \{(t_1, w_1), (t_2, w_2), \cdots, (t_m, w_m)\}$$

The dataset can be represented in the following flatten form:

$$\vec{d_1} = \{(egypt, 1/2), (protest, 1/2)\}$$
$$\vec{d_2} = \{(egypt, 1/4), (protest, 1/4), (time, 1/4), (presid, 1/4)\}$$

$$\vec{d_3} = \{(egypt, 1/3), (protest, 1/3), (presid, 1/3)\}$$
$$\vec{d_4}, \vec{d_5}, \vec{d_6}, \vec{d_7}, \vec{d_8} = \{(time, 1)\}$$
$$\vec{d_9}, \vec{d_{10}} = \{(real, 1/5), (time, 1/5), (oil, 1/5), (pric, 1/5), (obama, 1/5)\}$$

By using of terms in a pattern p, we can calculate pattern weight using $\vec{d}$:

$$weight(p, d) = \sum_{t \in p, (t,n) \in \vec{d}} n \tag{5.3}$$

The following demonstrates the pattern weight calculation for two patterns:

$$weight(< egypt, protest >, d_1) = 1/4 + 1/4 = 1/2$$
$$weight(< real, time, oil, pric >, d_9) = 1/5 + 1/5 + 1/5 + 1/5 = 4/5$$

For 1-pattern topic (i.e. term), its weight can be directly retrieved from the flatten vector:

$$weight(< egypt >, d_1) = 1/2$$
$$weight(< time >, d_2) = 1/4$$
$$weight(< time >, d_9) = 1/5$$

The updated weight is a more accurate representation for pattern p. In d_9, $weight(< time >)$ is significantly reduced and $weight(< real, time, oil, pric >)$ as a pattern is assigned with higher weight, appropriately representing its importance. Tweets can be represented in their pattern feature set with pattern weights (Table 5.4).

Tweet	Patterns Weight
d_1	<*egypt, protest*>$_1$
d_2	<*egypt, protest*>$_{1/2}$, <time>$_{1/4}$, <presid>$_{1/4}$
d_3	<*egypt, protest*>$_{2/3}$, <*presid*>$_{1/3}$
d_4, d_5, d_6, d_7, d_8	<*time*>$_1$
d_9, d_{10}	<*real, time, oil, pric*>$_{4/5}$, <*obama*>$_{1/5}$

Table 5.4: Tweets in pattern and weights representation

5.2.2 Adjusting Weight using Document Length

Although the weight of long patterns is now assigned appropriately according to their specificity, it is still unable to represent the topic weight in a dataset accurately. Table 5.5 shows the results of calculating topic weights using the weight distribution. Even though the pattern weights in a single document are distributed accordingly, the problem of short patterns overpowering long patterns still persists.

This problem is caused by generic term (<*time*>) occuring frequently, and mostly occurs in irrelevant tweets. By investigating these tweets, we found that most of them are short tweets and less likely to be related to news topics. They were not removed by the filtering in pre-processing step because they contain hashtag or mention, which confuses the classifier into considering them as relevant tweets. The length problem of tweets is studied previously by Naveed et al. (2011b), highlight that short tweets contain less useful information and less verbose. Low verbosity means information in long tweets is not repeated and therefore less redundant. This study also agrees with Zhao et al. (2010), who found that *"a single tweet usually covers a single topic."*. Tweet length can also be used to rank tweets (Nagmoti et al., 2010).

We use a similar idea here, to adjust the pattern weight by considering the length of

Topic	Total Weight
<*time*>	1/4 + 5(1) + 1/5 = 5.45
<*egypt, protest*>	1 + 1/2 + 2/3 = 2.17
<*real, time, oil, pric*>	2(4/5) = 1.6
<*obama*>	2(1/5) = 0.4
<*presid*>	1/4 + 1/3 = 0.58

Table 5.5: Topic weights using distribution

a tweet d, using a length factor (LF):

$$LF(d) = log(|\{t \in termset(d)\}|)$$

and the length adjusted weight of a pattern ($weight_l$) is calculated as:

$$weight_l(p, d) = weight(p) \times LF(d) \tag{5.4}$$

and a few examples of calculating $weight_l$ for patterns:

$$weight_l(< egypt, protest >, d_2) = 1/2 \times log(4) = 0.301$$
$$weight_l(< time >, d_4) = log(1) = 0.000$$
$$weight_l(< time >, d_9) = 1/5 \times log(5) = 0.140$$

The final weights of topics are calculated by aggregating the adjusted pattern weight (Table 5.6). With the pattern weights adjusted using length, the topic weight is more appropriately represented. The weight of <*time*> is significantly reduced, and the importance of long patterns <*real, time, oil, pric*> and <*egypt, protest*> is now higher than that of short patterns, reflecting their importance in a longer tweet.

Topic	Total Length Adjusted Weight
<*time*>	$1/4 \times log(4) + 5 \times log(1) + 1/5 \times log(5) = 0.290$
<*egypt, protest*>	$1 \times log(2) + 1/2 \times log(4) + 2/3 \times log(3) = 0.920$
<*real, time, oil, pric*>	$2(4/5 \times log(5)) = 1.118$
<*obama*>	$2(1/5 \times log(5)) = 0.279$
<*presid*>	$1/4 \times log(4) + 1/3 \times log(3) = 0.309$

Table 5.6: Final topic weights with length adjustment

5.3 Measuring Topic Popularity using Burstiness

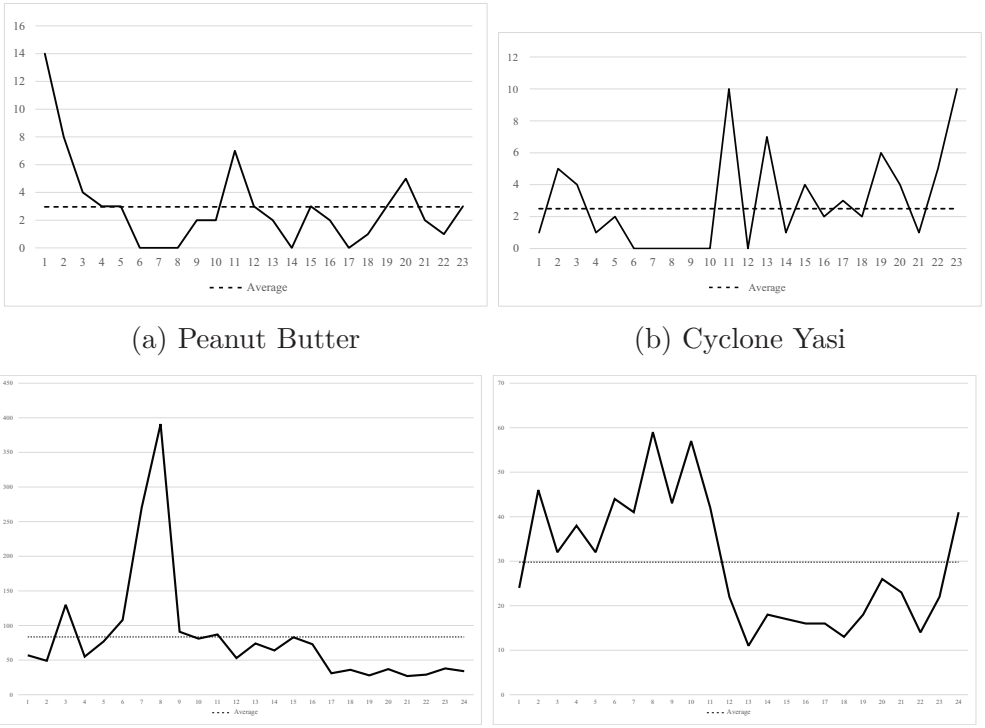
Computing pattern weights represents topic popularity from the volume perspective only. Compared with keywords, patterns are less likely to occur in high volume by chance, since they model terms relationship. This represents only one dimension of importance and is still insufficient for judging news relevance.

Topic weights are easily affected by popular users activity in Twitter. Popular figures such as celebrities (e.g. *"lady gaga"*, *"justin bieber"* and politicians (*e.g. "barack obama", "kevin rudd"*[1]) are always mentioned; common hashtags *#tgif* and *#nowplaying, #nowatching* are used repeatedly. Topic weight alone does not represent "*hotness*" of a topic, as high occurrence means only that a topic is always being discussed (Cataldi et al., 2010).

A topic is considered *'hot'* if its occurrences are significantly more frequently than average. One way to represent the *'hotness'* of a topic is to measure the burstiness. Burstiness captures the sudden spike that appears during a time series. News topics are a particular kind of topic displaying such *"bursty"* characteristics. Figure 5.2 shows the comparison between different topics. Using the concept of counting peaks, *Peanut butter* is less bursty since it has fewer peaks compared with *Cyclone Yasi*. From the volume perspective, *"Justin Bieber"*, despite having high tweets volume per hour, it has only one

[1]http://mashable.com/2010/05/14/twitter-improves-trending-topic-algorithm-bye-bye-bieber/

peak detected. In localized topics and recurring topics such as *#tgif, "good morning"*, we can often see this sort of recurring peaks especially when the tweets are contributed by users from different countries with timezone differences.



(a) Peanut Butter (b) Cyclone Yasi

(c) Justin Bieber (d) Egypt Protest

Figure 5.2: Burstiness comparion between bursty and non-bursty topic

Bursty topic modelling is influenced by the idea from Kleinberg (2003), which detects a sudden rising in the frequency of a stream. Kleinberg (2003) models an incoming stream using infinite-state automaton, and bursts are considered as state transitions. This approach does not suit the nature of tweets. Given the volume, it is not effective to perform such computation for all topics. Another burstiness calculation approach is to find peaks from a time series (Marcus et al., 2011). Peak counting scan through a time series to find peaks and the number of peaks indicates the importance; however, the number of peaks does not clearly indicate the amplitude of the peak over the time.

The idea proposed here is similar, but we want to quickly compute a localized numerical value that represents burstiness by using historical information from the last few observations. We calculate burstiness using the number of instances that are significantly higher than the historical averages. Tweets are evaluated using daily observation; the burstiness is measured every hour. If a topic has a significant increase of the volume of tweets above the average of the previous three hours, it is considered bursty. The burstiness ($Burst$) is calculated as:

$$Burst(p) = \frac{|coverset(p_j)|}{\sum_{k=j-3}^{k=j-1} |coverset(p_k)|} \tag{5.5}$$

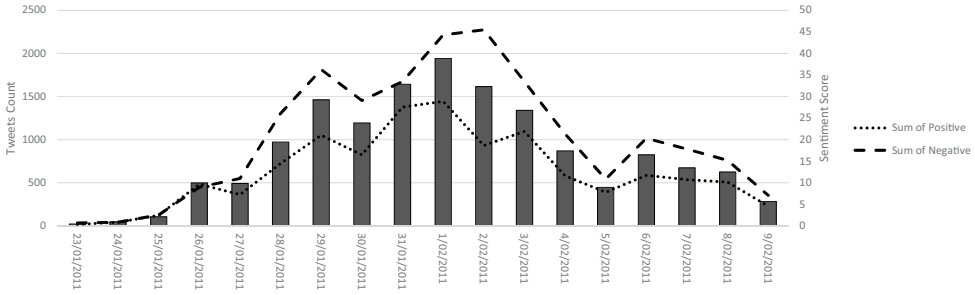
5.4 Measuring Interest Level using Sentiments

While burstiness can capture the hotness of a topic, it does not always present in all news event. Burstiness eliminates topics that display a flat characteristic, so overly popular topics are filtered. Occasionally, non-news related topics can trigger substantial message volume over specific time periods, to be mistaken as a bursty topic. This will lead to the failure of burstiness, and overuse of burstiness might also lead to low recall (Li et al., 2012; Ozdikis et al., 2012; Agarwal et al., 2012).

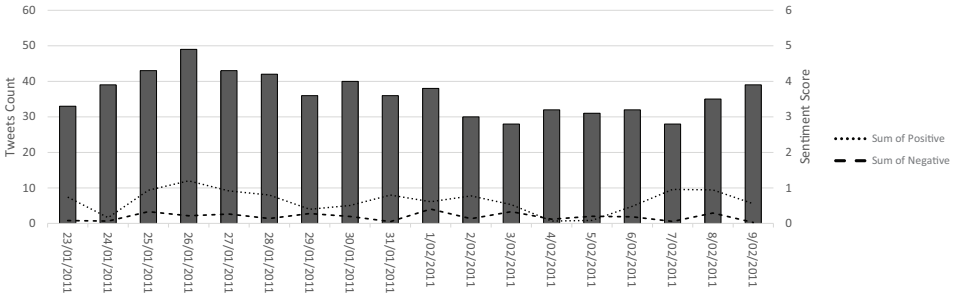
Another aspect of a news topic is the public interests. Controversial issues often trigger public interests and lead to user discussions with strong opinions. The public interest level can then be measured by detecting opinionated sentiment information. Sentiment is a feature often used to detect the public interest level in many microblog applications such as entertainment, politics and economics (Shamma et al., 2009; Bollen et al., 2011). We can utilize sentiments to measure the public interest level of a topic and to differentiate between topics that attract public attention and the boring ones .

Sentiment is also one of the key indicators in news topics (Kawai et al., 2007). As

seen in Figure 5.3, during the 2011 Egyptian riot, people actively participated in the discussion. A topic such as *"protest at tahrir square"* is likely to contain negative articles that will create sad sentiments for readers, while topic like *"release hostage in Cairo"* will create happy sentiments. A local topic *San Francisco*, is merely a ordinary topic where users discuss about local shops, or using location check-in services, therefore the level of sentiments is much lower. Users express their sentiments and opinions directly through tweets. The short nature of tweets even encourages readers to show stronger feeling while expressing sentiments about a topic (Bermingham and Smeaton, 2010).



(a) *Egyptian Protest*



(b) *San Fransisco*

Figure 5.3: Sentiment level for news related and non-related topics

Given the imposed length limit, a tweet can be considered as a sentence, and a shallow parsing method is sufficient for sentiment analysis (Bermingham and Smeaton, 2010). One technique that suits Twitter is the sentence level analysis. Our model adopts lexicon method to analyze the sentiments. Lexicon method is an unsupervised approach that analyze sentiments without manual supervision. It extracts opinionated terms and estimates the sentiment strength and orientation based on a dictionary, which

is constructed by human experts who affective content to lexicon tokens.

The Wilson lexicon list from the Multi-perspective Question Answering (MPQA) project (Wilson et al., 2005b) is selected as our corpus. Wilson lexicon list is the core of OpinionFinder (Wilson et al., 2005a), a sentiment analysis tool designed to determine opinions in the general domain. Wilson lexicon list provide a list of subjective words, each annotated with its degree of subjectivity (Table 5.7). Sentiment lexicons are annotated with different polarity and strength. We assign a score with a step of 0.5 for to indicate the difference between sentiment levels, similar to the settings used in the social media analysis study by Tjondronegoro et al. (2011).

Polarity	**Strength**	**Examples**	**Score**
Positive	Weak	*achievements, dreamy, interested, reputable, warm*	0.5
Positive	Strong	*awesome, breathtaking, charming, promising, wonderful*	1.0
Negative	Weak	*assault, bankrupt, breakdown, unlawful, weak.*	-0.5
Negative	Strong	*absurd, brutal, frustrated, horrible, remorse.*	-1.0

Table 5.7: Sentiment terms example from Wilson lexicon list and score

As shown in Table 5.8, the sentiment score s_d for a tweet d is calculated using the sentiment words w_i in the tweet:

$$s_d = \frac{\sum w_i}{|d|} \tag{5.6}$$

The example of tweets with sentiment score is shown in Table 5.8 and the aggregated sentiment score $Senti$ for topic p can be calculated using:

$$Senti(p) = \frac{\sum\limits_{d \in coverset(p)} s_d}{|coverset(p)|} \tag{5.7}$$

Tweet	s_d
Celebration for University of Manchester Nobel prize 2010 winners	0.272
BREAKING: Rahm Emanuel files motion to STAY Illinois Appellate Court ruling with Illinois Supreme Court. Will file appeal with court TUES.	0.214
David Cameron: "No Fears" Phone Was Hacked. Hmmmm funny that. Wonder why?	0.167
Egyptians Defiant as Military Does Little to Quash Protests	-0.350
Deadly Blast at Moscow's Main Airport Seen as Terror Attack	-0.228
Thousands Protest Against Jordanian Govt #egypt	-0.210

Table 5.8: Example of sentiment calculation

5.5 Twitter Properties

So far we have conisdered pattern weights, trending topics and public interest levels in our calculation. This information shows different aspects that affect news relevance, so now we consider unique actions that are used to spread news topics in Twitter. These properties have been shown to be useful in many studies, playing a significant role to indicate tweet importance, and to model and rank different events.

Determining news relevance topics in Twitter is challenged by noise, a well-known problem in tweets that severely increases the difficulty to judge news relevance. It has been reported that 14% of tweets are spam (Yardi et al., 2009) and only 3.6% of tweets are news related, which means we are dealing with with 96.4% of noise. In this context, irrelevant topics can be noisy terms such as *<shit>, <lol>* or just personal topic like *<happy birthdai>*.

5.5.1 Hashtags

Similar to most of the other social media tagging, a hashtag is a short label that is used to indicate the topic of a tweet for categorization and organization. Hashtags can also

present a collaborative view on news topics (Sankaranarayanan et al., 2009). For news related tweets, users actively make the effort to assign hashtags, as they see the value in doing, increasing the searchability of the tweets, and making contributions to the topic.

Hashtags follow a long-tail distribution in news topics. There is no governance of which hashtag to for different events, although sometimes during known events organizers will promote official hashtags such as *#emmys* for the Grammy awards, *#London2012* for the 2012 Olympics. However during crisis events, the hashtags (e.g. #eqnz, #qldflood) will be decided by the user communities.

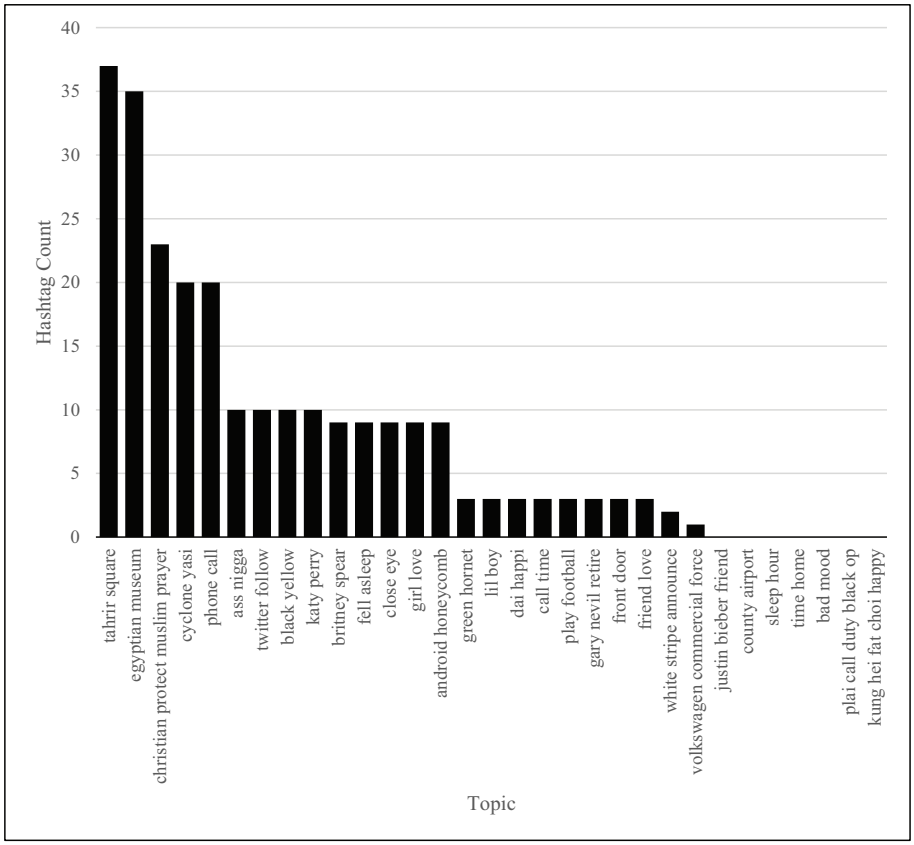


Figure 5.4: Hashtag count of topics

Figure 5.4 shows the number of hashtags used in various topics. Important topics such as *tahrir square*, *egyptian museum* will attract users to hashtag more frequently, and users are less likely to hashtag boring topics such as *sleep hour* and *bad mood*. Other problems we observed with hashtags include users tending to create hashtag randomly,

and the general uselessness of long hashtags. This is obvious: because of the length limit, users who genuinely want to spread news will opt for short and punchy hashtags, without wasting too much space.

Given that H_p represents the set of unique hashtags contained in p, the hashtag score HT can be calculated using:

$$HT(p) = \frac{|H_p|}{|coverset(p)|} \tag{5.8}$$

5.5.2 Urls

Twitter are primarily used to share and distribute information but its word limit restricts users to write verbose tweets with well-developed idea. Users then turn to include a url which directs their readers to external sources that include related materials such as blogs, videos and photos, to help other users to understand more of the tweet content. Popular urls receive more attentions and represent popular information (Cui et al., 2011).

For many events, related information is posted to Twitter in real time. For events that span across a long duration, such as floods that last for several days[2], or airport shootings that last for hours[3], users will continously post original url, adding to aspects of news and serving as evidence for news validity. The count of urls can then be used as an indicator to determine event popularity. These urls can be from news media, blog posts, or social images and video websites. Using url in tweets is also a good indicator for content interestingness. A study from Microsoft shows that by counting the tweets only is sufficient to be used as a feature, for detecting tweets that might be interested to wide audiences (Alonso et al., 2010).

Compared with hashtags, using url is simpler as most of the news relevant tweet

[2]http://www.smh.com.au/environment/weather/qldfloods

[3]http://www.news.com.au/world/gunman-paul-anthony-ciancia-shoots-one-dead-at-los-angeles-airports-terminal-3/story-fndir2ev-1226751686439

contain only one url, and multiple users can include the same url but not necessarily retweet each other (Galuba et al., 2010). A url can be considered as a good signal as it is language independent. The only issue while processing urls in tweets is that urls are always embedded in tweet using a shortening service, which turns a long url into a short version. These services can shorten links to under 10 to 20 characters. Some of these services are "http://tinyurl.com", "http://bit.ly", "http:// r.im", and "http://z.pe". For instance, *http://www.bbc.co.uk/news/world-africa-12289475* is shortened to *http://bbc.in/dT5AIM* using *bit.ly* url shortening service. The solution of this problem is to expand all shortened urls to its original form.

One key observation is that we want to measure the spread of urls in a topic. Some topics tends to be come from spammer or advertising topics, such as horoscope or weather forecast. Another type of redundancy is telling people to go to the same website. To address this redundancy, we measure only the unique urls from non-retweeted tweets, since retweet users can re-use the url. Let U_p be the collection of all unique urls extracted from $coverset(c)$, the url score can then be calculated using:

$$UR(p) = \frac{|U_p|}{|coverset(p)|} \tag{5.9}$$

5.5.3 Retweets

Retweet is the main mechanism to spread information within the Twitter network. Retweet is as simple as adding a "RT" prefix, or if the space allows; the user will also occasionally include their own content.

Retweet is an indicator of tweet popularity (Phuvipadawat and Murata, 2010), and also an indicator for tweet interestingness (Naveed et al., 2011a). This characteristic is extend to topic level. The main use of retweet is to amplify and spread tweets to new audiences. The number of retweets for a tweet is reflects its popularity, and its author's

popularity. From a topic point of view, if a topic contain more retweets, it shows that the topic is popular.

The retweet score RT of a topic c can then be calculated using the ratio of retweets to all tweets, using the following equation:

$$RT(c) = \frac{|d_r \in coverset(c)|}{|coverset(c)|} \tag{5.10}$$

This ratio intuitively represents the number of retweets contributing to the topic's popularity.

5.6 Topic News Relevance Scoring

In our model, we evaluate the topics that are relevant to news by judging the content and the usage patterns of Twitter activity. Other features such as following and follower, user verfication are not included, as they are mainly used to validate credibility (Castillo et al., 2011). Our focus is to identify topics that are news relevant and to inform users about emerging topics, where the credibility is beyond our scope of consideration. There are also incidents where verfied and credible user accounts can be exploited to disseminate false information[4].

We have now obtained various news relevance score measurements that describe the content (topic weight), context (burstiness, sentiments) and Twitter properties (hashtags, urls, retweets) represeting topic characteristics. We will need to train a model to compute the news relevance score using these measurements.

We apply *Logistic Regression* to learn the coefficients and to measure the impact of these features. Logistic regression is a probabilistic classification suitable for use when the dependent variable is binary. It is also used in other microblog tasks to determine

[4]http://edition.cnn.com/2013/04/23/tech/social-media/tweet-ripple-effect/

interestingness and popularity (Naveed et al., 2011a; Uysal and Croft, 2011). News relevance can be predicted as the likelihood score of a topic:

$$\begin{aligned} logit(p) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \\ prob(p) &= \frac{e^{logit}}{1 + e^{logit}} \end{aligned} \tag{5.11}$$

where we use the output from linear predictor function *logit* as the news relevance score, and *prob* represents the probability of a topic being news relevant.

Feature	**Coefficient**(β_i)
Topic Weight	0.735
Burstiness	0.829
Sentiment	-0.973
Hashtag	0.230
Url	1.790
Retweet	-1.832
Constant	-3.815

Table 5.9: Feature coefficients from logistic regression

Table Table 5.9 shows the coefficients produced from the logistic model. The model is trained using 1400 topics, manually selected from 14 days of tweets; each day contributes 100 topics classified using binary coding. Individual feature scores are summed linearly as the *logit* score, which represents the probability of a tweet being news related (Equation 5.11). It will be used to rank the news topics, and the top k topics are selected for evaluation.

5.7 Chapter Summary

This chapter presents a news relevance calculation algorithm to score topics extracted using PMM. The algorithm considers multiple factors that affects news relevance judgement.

As PMM treats patterns as the topic, we first address the topic weight issue by assigning proper weight to patterns in a document according to its importance, next we introduce a length factor, to provide pattern with appropriate weights according to tweet's length.

We then defined burstiness metric to capture trending topic and filter popular topics that are less meaningful. An unsupervised sentiment analysis is used to calculate a score to measure users' interest level of a topic.

Lastly, we calculate metrics based on three Twitter's activities that facilitate spreading of news topics. A logistic regression model is then train to classify tweets into binary news relevance, to learn the coefficient of individual features, which will be used to linearly calculate the probability of a topic related to news.

In next chapter, we will evaluate the performance of PMM. We test the robustness of pattern-based feature in retrieval tasks and shows PMM is able to detect news topics effectively and using multiple factors performs better than using single factor only.

Chapter 6

Evaluation

This chapter evaluates the performance of our news detection framework. First we evaluate performance of PMM in retrieval task to compare the difference between pattern-based model and term-based models. Then we evaluate the news detection algorithm using topics detected by PMM, and show the results of using single feature and combining multiple features. This chapter details the experiment design including the testing environment, datasets, baseline models, and evaluation methodology.

6.1 TREC Microblog Dataset

Obtaining a microblog dataset is a challenging task. Twitter TOS (terms of service) restricts the distribution of tweets in many forms. Most of the previous studies use proprietary datasets collected by individual researchers and are not available publicly, making it difficult for replicating their experiment results. TREC (Text REtrieval Conference) microblog dataset is the only public microblog dataset with annotations, which will be used for the evaluation.

The TREC microblog dataset (TRECMB) is the latest publicly available large-scale microblog dataset. TRECMB distributes only the unique ID of tweet. Each tweet needs to

be downloaded using Twitter REST API[1], in JSON (Javascript Object Notation) format, as shown in Figure 6.1. We pre-process the dataset using standard text pre-processing techniques and remove non-English tweets. The details of the final processed dataset are shown in Table 6.1.


```
"coordinates": null,
"created_at": "Sat Sep 10 22:23:38 +0000 2011",
"truncated": false,
"favorited": false,
"id_str": "112652479837110273",
"entities": {⊞···},
"in_reply_to_user_id_str": "783214",
"text": "@twitter meets @seepicturely at #tcdisrupt cc.@boscomonkey @episod http://t.co/6J2EgYM",
"contributors": null,
"id": 112652479837110270,
"retweet_count": 0,
"in_reply_to_status_id_str": null,
"geo": null,
"retweeted": false,
"possibly_sensitive": false,
"in_reply_to_user_id": 783214,
"place": null,
"source": "<a href=\"http://instagr.am\" rel=\"nofollow\">Instagram</a>",
"user": {⊞···},
"in_reply_to_screen_name": "twitter",
"in_reply_to_status_id": null
```


Figure 6.1: Example of JSON formatted tweet

Date	23rd January - 8th February
Total Tweets	16,141,812
Total Null Tweets	1,204,053
Total English Tweets	4,952,843
Total Retweets	2,596,642
Total Unique Tokens	27,240,636

Table 6.1: Details of TREC Microblog Track Dataset

TRECMB dataset contains 50 news relevant topics as shown in Table 6.2 (complete topic list is available at Appendix A), each of which is selected by the experts from National Institute of Standards and Technology (NIST) manually (Voorhees and Buckland, 2011), and with relevant assessment of tweets for each topic, which is reliable and robust for text model evaluations. (Soboroff and Robertson, 2003).

Figure 6.2 shows the example of a topic. Each topic contains a topic number and a title which describes the query. Individual topic also includes a *querytime* property which

[1]https://dev.twitter.com/docs/api

Topic	Query
MB001	BBC World Service Staff Cuts
MB010	Egyptian protesters attack museum
MB013	Oprah Winfrey half-sister
MB020	Taco Bell filling lawsuit
MB036	Moscow airport bombing

Table 6.2: Example of TRECMB topics

indicates the time when the query is issued and *querytweettime* property to indicate the maximum unique id of tweet when the query is issued. This is to limit the results and prevent the use of future information during experiments.

```
<top>
    <num> Number: MB001 </num>
    <title> BBC World Service staff cuts </title>
    <querytime> Tue Feb 08 12:30:27 +0000 2011 </querytime>
    <querytweettime> 34952194402811904 </querytweettime>
</top>
```

Figure 6.2: An XML formatted topic in TRECMB Dataset

6.2 Evaluation Metrics

Evaluating the performance of a microblog system is challenging requires the processing of large amount of short-fragments texts, which often leads to external information in different formats such as text, video, audio, websites, news articles and blog posts. This is difficult to standardize and process by machines automatically, for its complexity as discussed in the literatures and shown in Figure 6.3.

There are a high number of tweets retrieved for each topic and relatively fewer relevant tweets in this retrieval contains. About 800 tweets are retrieved for each topic, but on average only 50 tweets per topic are relevant (i.e. 92% of retrieved tweets are irrelevant). There are 26 topics which contains less than 50 relevant tweets and only 10 topics contains over 100 relevant tweets. Detail statistics of TRECMB relevant tweets are shown in Table

Tweet	Score
BBC World Service | Radio Frequency Guide [del.icio.us]: RT @BBCPhilippaT: Please RT @BBCNewshour: SW frequencie... http://bit.ly/ihH5Ie	-2
World News: US diplomat kills two Pakistanis: An American diplomat in the Pakistani city of Lahore has shot and ... http://bbc.in/f8Ae4H	0
Sad. I went through that myself in the first round 5 years ago: BBC World Service to cut up to 650 jobs http://bit.ly/dGkyzp -	1
BBC News - BBC World Service cuts to be outlined to staff http://www.bbc.co.uk/news/entertainment-arts-12283356	2

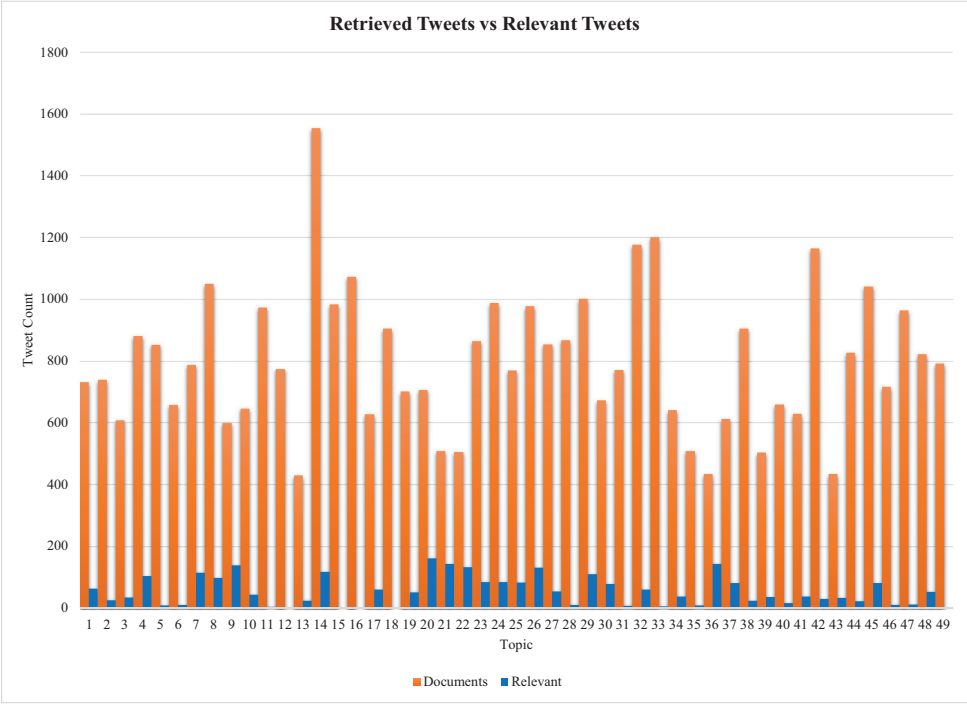
Table 6.3: Examples of tweet relevance score for Topic MB001

6.4. A relevance score is assigned for each retrieved tweet in a topic in TRECMB. The score is assigned by human experts manually, each with a relevance score between -2 and 2, as shown in examples from Table 6.3.

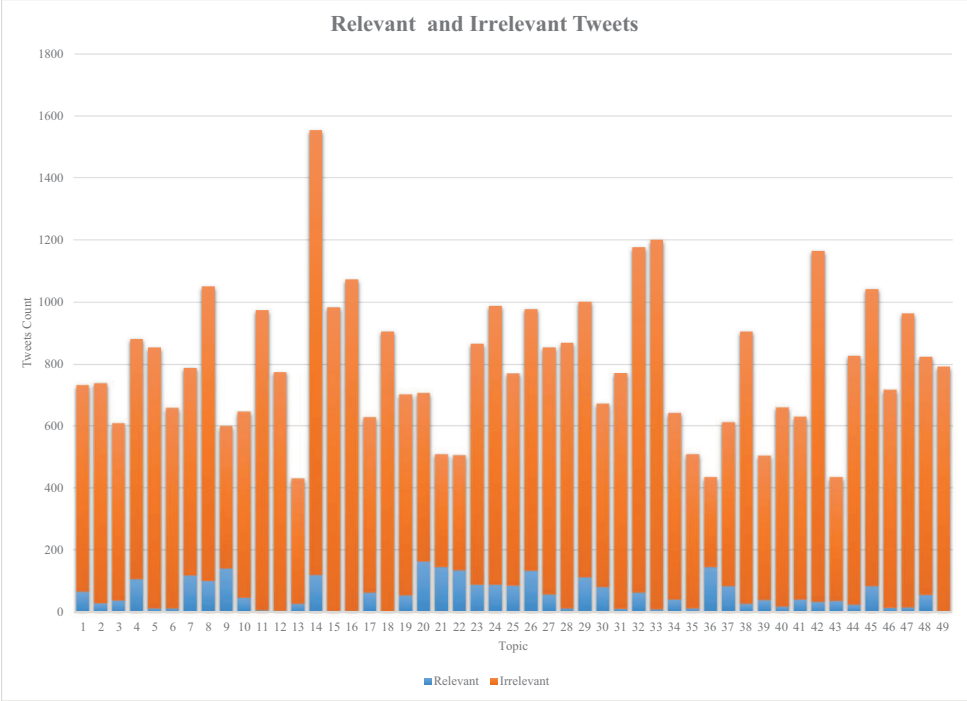
ID	#D	#R	ID	#D	#R	ID	#D	#R	ID	#D	#R	ID	#D	#R
001	731	65	011	975	6	021	509	145	031	771	10	041	629	40
002	738	28	012	776	4	022	506	134	032	1177	62	042	1165	32
003	608	37	013	430	27	023	867	87	033	1201	8	043	435	36
004	883	106	014	1553	119	024	989	87	034	641	40	044	829	24
005	855	11	015	985	2	025	769	84	035	509	11	045	1043	83
006	658	12	016	1074	2	026	979	132	036	435	145	046	716	13
007	790	117	017	628	62	027	856	56	037	612	83	047	965	14
008	1051	99	018	907	1	028	870	12	038	907	26	048	825	55
009	599	140	019	701	54	029	1003	111	039	504	38	049	794	2
010	646	46	020	706	163	030	672	80	040	659	18	050	NA	NA

Table 6.4: Number of Relevant Tweets(#r) and total number of Retrieved Tweets (#d) by accessor in TRECMB

Two standard IR measurements: precision and recall are used for the evaluation. Precision is the fraction of retrieved tweets that are relevant to the topic; recall is the fraction of relevant documents that have been retrieved. Precision and recall can be calculated using the following formulas:



(a) Retrieved tweets and irrelevant tweets



(b) Relevant tweets and irrelevant tweets

Figure 6.3: Tweets relevance assessment for topics in TRECMB dataset

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

The judgements can be represented in binary form according to the definitions (Table 6.5). TP (True positive) denotes the number of tweets the system correctly identify as relevant, FP (False Positive) is the number of tweets the system identifies as relevant. FN is the number of relevant tweets the system fails to identify and TN (True Negative) is the number of tweets that the system correctly identifies as irrelevant.

		Human Judgement	
		yes	*no*
System Judgement	*yes*	TP	FP
	no	FN	TN

Table 6.5: Contigency Table

In tweets retrieval, it is unfair to judge the performance of a model by evaluating it with only a few topics; it is more reasonable to judge the overall performance of a system. In addition, it is hard to evaluate the recall for entire corpus. Therefore we evaluate only the top k tweets returned. This is reasonable since most users will focus on the first set of results in most cases, especially when they have found the information they required. We follow the guidelines in TREC Microblog Track and evaluate precision using $k = 30$ results returned (P@30), and a Recall level precision ($R-PREC$). Mean Average Precision (MAP) is used to represent the effectiveness of a model across the collection of topics, defined as:

$$MAP(Q) = \frac{1}{|Q|}\sum_{j=1}^{|Q|}\frac{1}{m_j}\sum_{k=1}^{m_j} Precision(R_{jk}) \tag{6.1}$$

R-Precision denotes the precision after R documents have been retrieved (Table 6.6). The average R-Precision is computed by taking mean of R-Precisions of the model. Recall level precision reduces the rank effect if the number of retrieved documents is less than the required.

Document Level	**Average**
At 5 docs	0.4280
At 10 docs	0.3960
At 15 docs	0.3493
At 20 docs	0.3370
At 30 docs	0.3100
R-Prec	0.3640

Table 6.6: Example of R-Precision

6.3 Baseline Models and Settings

We compare the retrieval performance of PMM with other baseline models including term-based models and pattern-based models with different weighting strategies.

6.3.1 Term Based Model

Term Frequency Inverse Document Frequency

Term Frequency Inverse Document Frequency (TF-IDF) is a classic model widely used in

many IR applications. It has also been used in various microblog studies for its maturity and sensitive to word changes. Inverse document frequency (IDF) is used particularly to reduce the effect of document length and to filter off irrelevant terms that appear frequently. In TFIDF, given a query q and a document d, their similarity can be measured using cosine similarity:

$$Cosine(q,d) = \frac{\sum_w weight(w,q) * weight(w,d)}{\sqrt{\sum_w weight(w,q)^2}\sqrt{\sum_w weight(w,d)^2}} \tag{6.2}$$

Lucene

Lucene is a state-of-the-art high performance, full-featured text search engine which is used by TREC organizers to create the baseline. It has been also used in many IR studies and commonly used as a baseline in many tasks of Text REtrieval Conference (TREC).

6.3.2 Pattern Based Models

As our model is based on the pattern model, it is important to compare and contrast the performance between PMM and the other pattern-based models. The main aim is to investigate the effect of using sequential patterns and non-sequential patterns. The minimum supportmin_sup is set to 0.2, which a pattern is considered frequent if it appears in n tweets of total m tweets where $n/m \geq 0.2$, to eliminate noisy patterns.

Non-Sequential Closed Frequent Pattern

The Non-Sequential Closed Frequent Pattern (NSCPM) is one of the basic pattern-based methods that find patterns without considering the order of terms. For instance, $\{t_1, t_2, t_3\}$, $\{t_1, t_3, t_2\}$ and $\{t_3, t_2, t_1\}$ are considered the same, and the support will be calculated for one pattern only. This is the primary baseline for pattern model .

Term Weights using Pattern Deployment

In Pattern Taxonomy Model (PTM), the weight of a term can be derived by its appearance in frequent patterns (Wu et al., 2006). That is to say, term t_i is considered more important than term t_j, if t_i appears in more frequent patterns. For example, given frequent patterns <*apple, ipad*> and <*apple, ipod*>, the support of *apple* is then equal to 2 where *ipod* and *ipad* each equal to 1. Term weight $w(t)$ can then be calculated by aggregating the support of frequent patterns where t appears:

$$w(t) = \sum_{i=1}^{|P|} |\{p|p \in P_i, t \in p\}| \tag{6.3}$$

Pattern Weights using Terms

Another strategy to weigh the importance of a pattern is to use the weights of terms in the pattern. A pattern p_i is more important than p_j if the aggregated weight of $t \in p_i$ is higher than the aggregated weight of $t \in p_j$. The weight $w(p)$ can be calculated by aggregating the term frequency for all terms in p.

$$w(p) = \sum_{t \in p} w(t) \tag{6.4}$$

Pattern and Terms with Deployed Weight

Pattern and terms with deployed weight is to combine the weights of terms and pattern without bias towards any of the features.

6.4 Evaluation of Pattern Model for Microblogs (PMM)

This evaluation aims to answer whether PMM is a suitable model for microblogs and to compare its performance with other existing models. Therefore, we apply PMM at a practical microblog retrieval task, the TREC 2011 "real-time" search task.

6.4.1 Query Expansion

One long discussed issue in IR tasks is the query mismatch problem, where the query terms provided by users are not the actual terms used in the result documents. This problem is caused by the short query, as users are unable to specify the context of the query terms. This problem become worse as the query terms and result documents (i.e. tweets) are both short. The short length of Twitter has always been an issue affecting the search performance using existing algorithms. One solution to this problem is to expand the initial query, by including more related terms and aiming to retrieve more relevant results (Suh et al., 2010).

Microblog queries are always too short to accurately describe actual user intention. On average, a microblog search contain 2.8 words; a web search contains 3.75 (Teevan et al., 2011). Different users apply different terms to describe similar topics and concepts, which makes the retrieval task even harder and produces inconsistent retrieval results. Terms frequency in a tweet is insignificant to provide distinction in meaning and the sparsity negatively impacts the retrieval performance. A potentially relevant tweet will not be retrieved at all if among its few terms it does not contain any of the query terms. This problem become more significant for microblog compared with full length documents, as word limit of tweets prevents authors from elaborating content verbosely (Teevan et al., 2011).

Query expansion (QE) is widely used in Information Retrieval to address the short

query problem. The aim of query expansion is to enhance the initial query by including more relevant terms. The five main steps of a typical QE are: retrieve initial result set using original query, select top k results using relevance feedback, extract representative terms from the top k result set, form a new query by expanding the original query using these terms, and perform another query with the new query. The expanded query is then run again to obtain to final result set.

Query expansion techniques can be classified into two types based on the information source: global analysis and local analysis. Global analysis utilizes corpus-wide statistics such as term co-occurrences, to expand the results using terms pair with the highest similarity. Performance of global analysis is generally robust in static document collection, but it requires a considerable amount of resources when the data collection is huge. Therefore it is not suitable for fast-changing dataset such as tweets.

Different from global analysis, local analysis requires only some initially retrieved documents to expand the query. One well-known local analysis technique is relevance feedback, which reformulates the query based on retrieved document relevance. This approach achieves reasonably good performance when sufficient positive relevance judgement is provided, but always requires manual efforts.

To overcome the requirement of manual judgement, the pseudo-relevance feedback (PRF) approach is introduced. PRF blindly assumes the top k ranked results is relevant and extracts expansion terms from these documents only. This technique also overcomes the problem of vocabulary mismatch between query and relevant document by broadening the scope of query while staying on topic (Manning et al., 2008). To avoid over expanding the query, k is also kept small (e.g. $k = 5$) for optimized retrieval performance (Metzler and Croft, 2004). The benefit of PRF is that it improves retrieval performance without external interaction (Manning et al., 2008), which is ideal for tweets retrieval since it is unrealistic for humans to perform multiple feedback iterations.

Query expansion has been shown useful in improving the text documents retrieval in various models (Zhai and Lafferty, 2001; Algarni et al., 2009; Massoudi et al., 2011). In a study by Massoudi et al. (2011), query expansion is used together with other Twitter features as quality indicators to improve search results. The key aspect of our query expansion approach is that expansion is done over a stream of short and noisy messages using localized analysis without any external knowledge.

Here we formally define the query expansion technique used in our study. The flow of the technique is logically controlled a main control process (Algorithm 5).

Algorithm 5: Main Control Process

Input
- A set of tweets. $D = \{d_1, d_2, d_3, ...d_n\}$
- Initial query, q

Method
1. Use q to retrieve top 100 tweets in D.
2. Sort the retrieved tweets based on time.
3. Form training set Ω using top 10 of the sorted tweets.
4. Form expanded query Q using terms from Ω.
5. Use Q to retrieve 1000 tweets from D and sort based on time.
6. Use top 30 tweets as the final result.

Given an initial query q, we retrieve a set of tweets ranked by its similarity. The top k ($k = 100$) tweets are then ordered by time in reverse order for selecting top 5 tweets (examples in Table 6.7) to form the training set Ω. All the terms in Ω are then used to expand q to form Q Table 6.8. Expanded query Q is then used to perform another query to select top 1000 relevant tweets. These tweets are again ordered by time in reverse order and the top 30 tweets are selected as the final results.

Tweet	Tokens
The Egyptian authorities may impose curfew in the coming hours according to Al Jazeera	egyptian, authorities, impose, curfew, coming, hours, according, jazeera

In Pictures: Revolt in the Nile: Images of the thousands of Egyptian protesters that defied a curfew all over Egypt.	pictures, revolt, nile, images, egyptian, protesters, defied, curfew, egypt
Sign of positive change: Unlike many Egyptian 'fathers', Mubarak has imposed his curfew fairly upon both girls AND boys	sign, positive, change, egpytian, fathers, mubarak, curefew, fairly
The curfew which the Egyptian army declared in Cairo is pretty good: One million !!! in Tahrir Square. 19 millions at home.	curfew, egyptian, army, declared, cairo, pretty, milllion, tahrir, square
Middle East unrest: Egypt imposes night curfew after day of riots: Egyptian state TV says President Hosni Mubarak	middle, east, unrest, egypt, imposes, night, curfew, riots, egyptian, state, president, hosni

Table 6.7: Examples of top tweets from initial query

Query: egyptian curfew
egypt jazeera arab tahrir mubarak military revolution protesters government january muslim police

Table 6.8: Query expansion result

6.4.2 Tweets Ranking

The final step in our retrieval model is a ranking mechanism that returns a list of tweets based on the user query. In social media, systems often ignore content relevance and assume that users are interested in latest information only(Mishne et al., 2006). For instance, search results provided by Twitter.com are ranked in reverse chronological order, which provides no guarantee that the most interesting tweets will be on top (Nagmoti et al., 2010) This does not help users to retrieve information: what users generally want is to get up to speed with the topic they search for (Soboroff et al., 2012).

Properties of social media can be used to improved retrieval performance (Mishne, 2007). As the scope of our research is limited within the context of news, the following Twitter specific features are considered, based on the requirements and needs for news retrieval.

- **Hashtags**: The proportion of hashtags in tweets.

$$H(d) = \frac{\text{total no. of hashtags}}{\text{total number of terms}}$$

- **Url**: Indicates the presence of URL containing video, picture, or articles.

$$U(d) = \begin{cases} 1, & \text{if } d \text{ contains URL} \\ 0, & \text{otherwise} \end{cases} \tag{6.5}$$

- **Retweet** : A binary value indicates a retweet.

$$R(d) = \begin{cases} 1, & \text{if } d \text{ is a retweet} \\ 0, & \text{otherwise} \end{cases} \tag{6.6}$$

6.4.3 Similarity Metric

In vector space models, queries and documents are represented as vectors in dimensional space V, where V contains all distinct terms in the document collection. Document similarity is determined by the similarity of their content vector. This has led to three problems: (i) low frequency terms in the collection will be assigned with relatively high weight; (ii) similarity score might be low due to the absence of terms in query, and (iii) semantically related terms that do not appear in query will not be retrieved. Therefore, cosine similarity will not be appropriate for use as the similarity measure, instead, the

Jaccard Index is more suitable. Equation 6.7 defines the Jaccard similarity function between a query q and a tweet d:

$$Sim(q,d) = \frac{|terms(q) \cap terms(d)|}{|terms(q) \cup terms(d)|} \tag{6.7}$$

where the *terms* function is to extract the termset from a query or a document.

Jaccard similarity is used as the similarity function in our approach. Classic cosine similarity (Equation 6.2) measures the similarity between two tweets by angle in the vector space. In our case, the tweet is represented using patterns and the expanded query; term weights are insignificant as the difference between terms is subtle. Furthermore, terms that do not co-occurred between query and tweet will further degrade the performance. The key in our model is to measure the overlapping of patterns between the query and tweet, therefore, Jaccard index is more appropriate.

6.4.4 TREC Microblog Track Retrieval Results

Figure 6.4 shows the overall results between models and detail precision at different recall levels are listed in Table 6.9. The complete results of TREC 2011 Microblogs is available at Appendix B.1 and topic listing at Appendix A. Performance of individual topic compared with the median precision provided by TREC are shown in Figure 6.5.

Note that non-sequential frequent pattern (NSCP) based methods struggle to perform in different settings and weightings. This is due to large amounts of noise generated during the pattern generation process. Another key finding is that Sequential Closed Pattern (SCP) based methods perform fundamentally better than any of the pattern-based model. In addition, cosine similarity is found to be not suitable for pattern based approaches, as the additional noises generated during pattern mining leads to a performance degradation.

Performance of TF-IDF is acceptable at a lower recall level, but NSCP methods

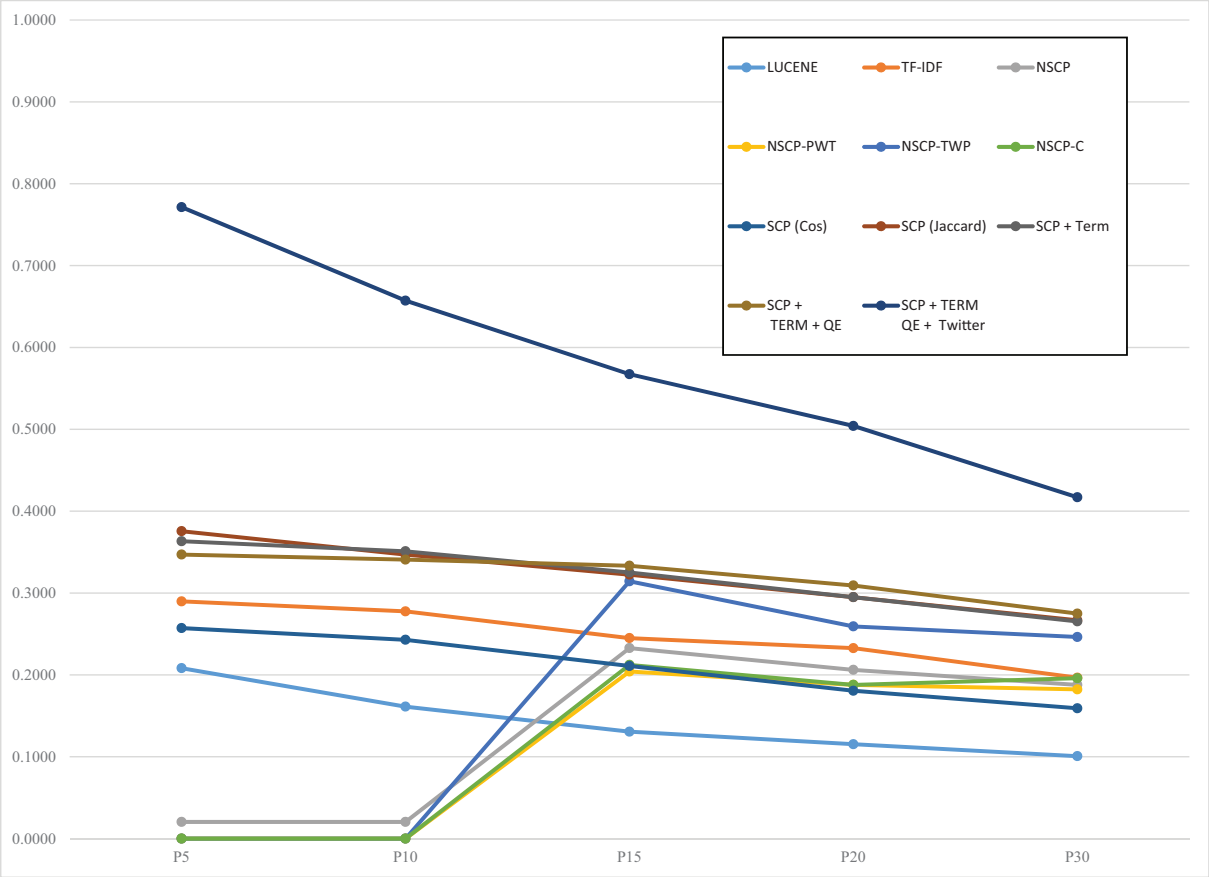


Figure 6.4: Average precision @ k for all models

are able to obtain more positive results after P10. On the recommended P30 level, SCP approaches outperform all methods. Query expansion (QE) does not improve the performance on the overall system level, but introducing Twitter features helps to retrieve many more related results.

MB002 (2022 Fifa soccer) is the topic which all models struggle to retrieve relevant results, due to the stopword removal which causes term *"2022"* not to be included. Other poorly performed topics includes MB014 (release of The Rite) and MB015 (Thorpe return in 2012 Olympics) for the same reason.

SCP based approaches effectively capture patterns contain named entities, which has led to good performance for topics such as the MB001 using pattern <*bbc, world, service*>. SCP based approaches are unable to handle topics that contain no meaningful patterns,

	Mean Average Precision (MAP)					
	P5	**P10**	**P15**	**P20**	**P30**	**R-PREC**
LUCENE	0.2082	0.1612	0.1306	0.1153	0.1007	0.1432
TF-IDF	0.2898	0.2776	0.2449	0.2327	0.1966	0.2483
NSCP	0.0204	0.0204	0.2327	0.2061	0.1878	0.1859
NSCP-PWT	0.0000	0.0000	0.2041	0.1878	0.1823	0.1848
NSCP-TWP	0.0000	0.0000	0.3143	0.2592	0.2463	0.2512
NSCP-C	0.0000	0.0000	0.2122	0.1878	0.1959	0.1978
SCP (Cos)	0.2571	0.2429	0.2109	0.1806	0.1592	0.2101
SCP (Jaccard)	0.3755	0.3469	0.3224	0.2949	0.2667	0.3213
SCP + Term	0.3633	0.3510	0.3252	0.2949	0.2653	0.3199
SCP + Term + QE	0.3469	0.3408	0.3333	0.3092	0.2748	0.3210
SCP + Term + QE + Twitter	0.7714	0.6571	0.5673	0.5041	0.4170	0.5251

Table 6.9: Performance comparison for all models

or term order is not important, such as MB011 (Kubica Crash). This is because the meaning of query will not be changed even if the order between *kubica* and *crash* has changed. On the average level of MAP and R-PREC, query expansion does not seem to be improving the retrieval performance significantly. But in reality, this is not the case. For instance in MB007, query expansion improves the original result by 1.6 times from 0.2933 to 0.7667. More related terms (*officer, charges, consular, shoot*) are added to the original query and leads to a much higher precision.

On topics that are prone to be noisy such as entertainment (MB013), query expansion causes poor performance as more noisy terms are appended to the queries. In MB013, the performance dropped from 0.6000 to 0.2500 for almost 60%, due to many irrelevant terms such as *family, prompt* and *buzz*, are appended to the query. In MB030, even with useful pattern <*keith, olbermann*> in the topic, since the tweet results are generally short, any additional terms appended to the query will adversely affect the retrieval performance.

Lastly, features from Twitter have been seen to be able to improve retrieval results

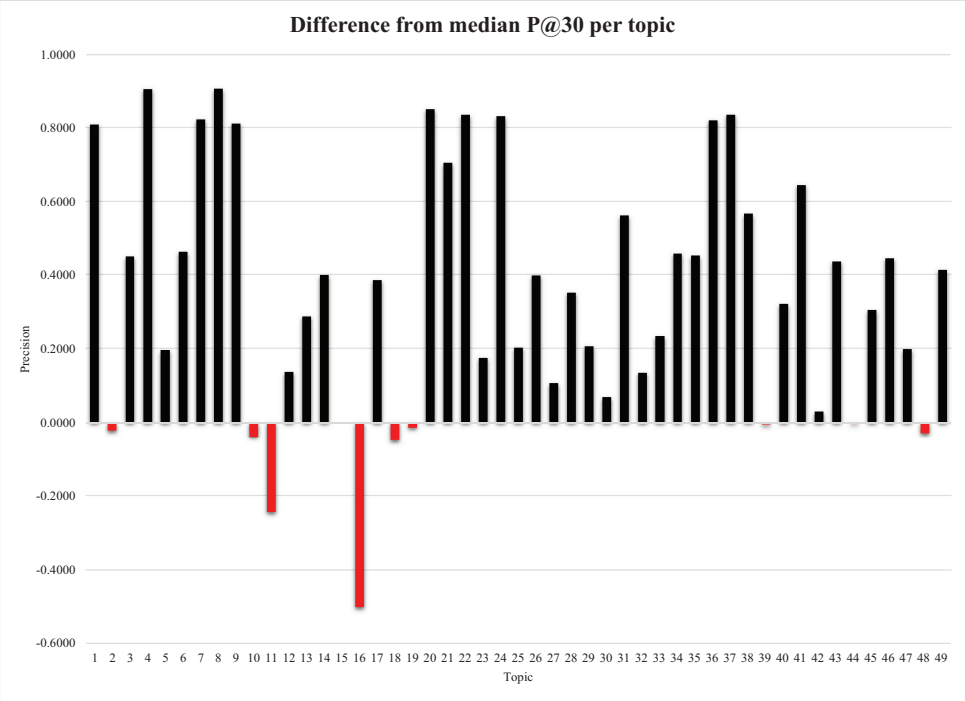


Figure 6.5: Difference with TRECMB median precision

significantly. This result shows that news relevant tweets usually contain hashtags and urls that provides more information. The cases where Twitter features do not improve the performance signficantly or fail to perform are MB015 (*Thorpe return in 2012 Olympics*), MB016 (*release of Known and Unknown*), MB019 (*Cuomo Budget Cuts*) and MB027 (*reduce energy consumption*). In particular, MB027 is a query that is challenging for retrieval, as the topic is more of a generic topic similar to MB029 (*global warming and weather*). However, Twitter features are reliable for use in improving tweets retrieval overall.

6.5 News Topics Evaluation

The evaluation process of our news topic detection algorithm includes baseline model selection, manual assessment and the result evaluations.

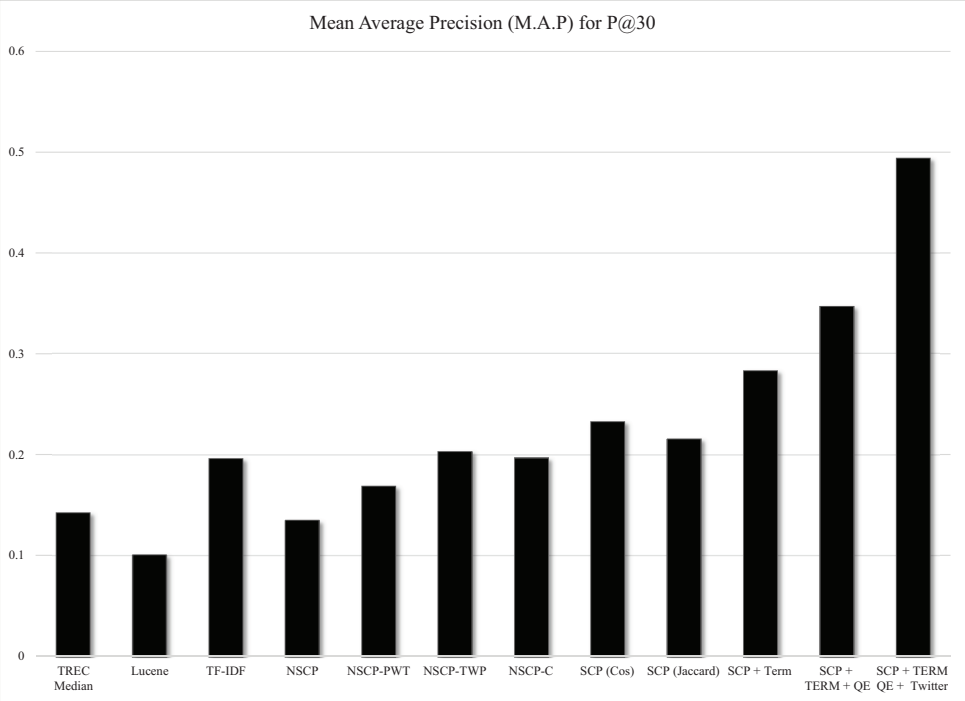


Figure 6.6: Performance comparison of retrieval models

6.5.1 Baseline Models

Baseline model selection involves identifying suitable systems and models to serve as the performance benchmark for our algorithm. Previous approaches focus primarily on detecting events (Agarwal et al., 2012; Cui et al., 2012; Gayo-Avello et al., 2011; Gupta et al., 2012) and trends (Benhardus and Kalita, 2013; Goorha and Ungar, 2010; Mathioudakis and Koudas, 2010), but not news. Available news detection systems either require manual input, or do not focusing on news detection. For instance, TwitterStand (Sankaranarayanan et al., 2009) requires a list of trustworthy news sources in order to work; TwitInfo (Marcus et al., 2011) and Eddi (Bernstein et al., 2010) allow users to browse news events, but they focus on visualization and user defined events, which are different from our goals.

We further investigate the text mining and topic detection technique, K-Means. K-means is a clustering technique that uses term features to group documents according to their similarity. However, K-Means requires number of topics k to be pre-determined, but

Model	Description
Lucene	Lucene Baseline provided by TREC
TF-IDF	Term Frequency Inverse Document Frequency
NSCP	Non-sequential Closed Frequent Pattern
NSCP-PWT	NSCP (Pattern Weighted Terms)
NSCP-TWP	NSCP (Term Weighted Pattern)
NSCP-C	Combine NSCP-PWT and NSCP-TWP
SCP	Sequential Closed Frequent Pattern
SCP + Term	Sequential Closed Frequent Pattern and Term
SCP + Term + QE	SCP + Term + Query Expansion
SCP + Term + QE + Twitter	SCP + Term + QE with Twitter Features

Table 6.10: Summary of Models

this is hard to estimate from a large and constantly changing dataset such as Twitter.

Another concern is to use topic models such as LDA (Blei et al., 2003). Topic modelling is a type of unsupervised probabilistic modeling technique that has been used in many topic detection tasks. One particular model is the scalable version of LDA (Hoffman et al., 2010), which uses single pass document collection and able to process tweets in an efficient online manner. Although suitable for use in Twitter environment for detecting topics or events, this model also requires number of topics k to be pre-determined. The k in both K-Means and Online-LDA is a static number, which is different from our objective. We aim to find as many topics as possible, since the number of news topics is always unknown.

Other models require manual effort: for instance, Labelled LDA (Ramage et al., 2010) requires a set of manually classified tweets in order to work. An event detection model by Becker et al. (2011) requires a classification model to be trained. In another model by Diao et al. (2012), users profile are required to select training topics; an emerging topic detection model by Cataldi et al. (2010) requires supervised term selection.

Due to the lack of formal baselines that evaluate news detection using public datasets, we follow the baseline selection approach in the study by Cataldi et al. (2010). We

TopicID	Topic Title
MB002	2022 FIFA soccer
MB005	NIST computer security
MB011	Kubica crash
MB014	release of "The Rite"
MB015	Thorpe return in 2012 Olympics
MB016	release of "Known and Unknown"
MB018	William and Kate fax save-the-date
MB030	Keith Olbermann new job
MB038	protests in Jordan
MB048	Egyptian evacuation

Table 6.11: Topics with poor performance

present the evaluation results of using individual features, including pattern weights, temporal, sentiments, Twitter features, compared with multiple features model using logistic regression. This evaluates the performance between using single feature and combining multiple features.

6.5.2 Manual Assessment

Three assessors are employed for the evaluation process. They are Twitter power users who are experienced in journalistic activities, use Twitter regularly to explore news contents and are familiar with various usage syntaxes.

The assessors are presented with the topics and their related tweets. Tweets are presented in reverse time order. The assessors need to decide if a topic has news relevance using binary coding. News relevance can be decided with the help of media coverage. A topic is considered as news related, if it is reported by at least two media outlets. This idea implicitly derives news relevance based on the judgement of news editors who are professionally trained to judge news relevance.

For news topics that are not reported by media, assessors label each topic according to their judgement. Assessors are encouraged to utilize external information such as urls in tweets, or other tools such as Google and Wikipedia, to assist them in judging news relevance. A topic is considered news relevant with the agreement of at least two assessors.

6.5.3 Results of Baseline Models

Figure 6.7 show the performance comparison between PMM and other baseline models using precision @ k, using the TRECMB dataset. Result shows that single features perform reasonably well and that the performance is significantly improved when multiple features are combined.

One key consideration is the Part-of-speech (POS), where we exclude single patterns that do not contribute much to news topics, such as adjectives (e.g. *hot, cold, large*), adverbs (e.g. *carefully, mainly*), and verbs (e.g. *run, sleep, eat*). Pronouns (e.g. *I, he, she, it*) and articles (e.g. *the, a*) are eliminated by stopword removal during pre-processing. These patterns dominates the top positions with high numbers of counts, and also detects meaningless topics (e.g. <*love*> <*hate*>, <*like*>, <*support*>). The support count for these topics can be up to 100 times more than for usual topics, adversely affecting the news detection performance of all features. Here we are focusing on noun tags and proper noun tags such as person, locations and events, as they are key subjects in news topics.

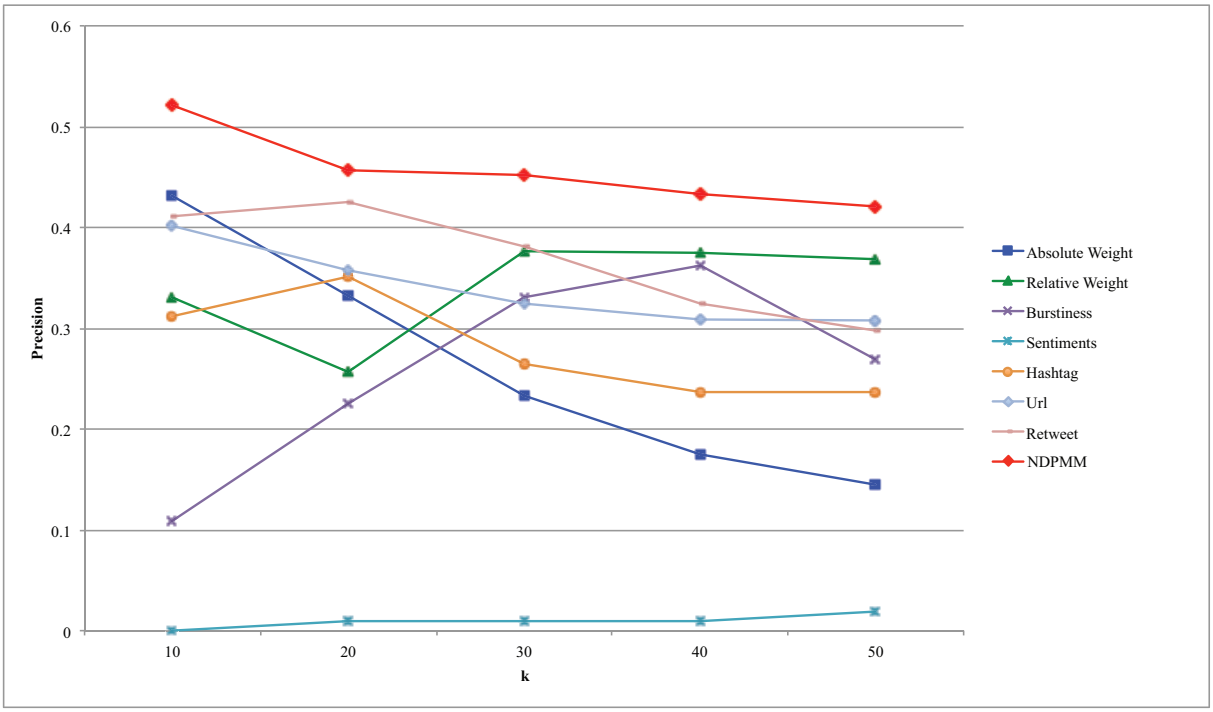


Figure 6.7: Performance comparison between PMM and baseline models

Experiment results show that rich amounts of important news topics can be detected. Particularly, most of the topics are of entertainment, politics and technology, and most of the news are from United States of America (USA). This could be due to the majority of Twitter users are from the United States of America. Entertainment topics are popular as many celebrities actively maintain their social media presence by regularly posting tweets about their updates and activities. They also have a significantly large number of followers who are also actively tweeting and retweeting their activities.

Major topics detected include the uprising of Egypt Revolution—where demonstrations, riots and protests took place to demand the overthrow of President Hosni Mubarak (Ahmed, 2011). Twitter played a crucial role in this incident, as access to other major social media were completely restricted except Twitter. This incident also unleashed the potential use of Twitter, such as using hashtag *#jan25* to label tweets; using retweets and mentions to spread and communicate information; sharing urls and multimedia content to provide footage of areas that are inaccessible to mainstream media. This is a noteworthy

incident that highlights the remarkable use of social media to coordinate, record and disseminate information during a critical situation.

Other news topics in the dataset include Rahm Emanuel—former Obama Government's Chief of Staff Rahm Emanuel, who ran for and was eventually elected as the first Jewish Mayor for Chicago (Editor CNN, 2011); Moscow Domodedovo Airport—Russia's busiest airport, which was attacked by terrorist, causing many casualties (CNN, 2011). More examples of news topics of other categories are shown in Table 6.12.

Topic	Description	Reference
<*charli sheen*>	Actor Charlie Sheen admitted into hospital	(Allen, 2011)
<*oprah half sister*>	Celebrity host Oprah Winfrey found out she has a secret sister	(Gabbatt, 2011)
<*mandela admitted hospital*>	South African's Ex-President Nelson Mandela was admitted into Johannesburg's hospital	(Smith, 2011)
<*all star game*>	Coverage of 2011 NBA All Star Game	(USA Today, 2011)
<*julian assang defend wikileak*>	Founder of Wikileaks - Julian Assage defend himself on a TV show	(Parr, 2011)

Table 6.12: Examples of news topics in TREC Microblog Dataset

Absolute Weight

Absolute weight represents topic importance by merely counting the pattern occurrence, without considering other factors such as pattern weight and length. As expected, absolute weight captures topics with high tweet volume (Table 6.13), such as <*justin bieber*>, <*lady gaga*>, <*obama*> and <*presid obama*>. These topics, although non-spam and written by genuine users, are neither contributing knowledge to any event nor reflecting any real world news topics.

Tweets
Hell yeah.. These girls love him to death. RT @iDavidaG_84: #iBet Justin Bieber's movie is gonna sell out
Selena Gomez will be attending Justin Bieber "Never Say Never" Movie Premiere today!! excited to see what she wears!
Nicola Formichetti said that "God Only Knows" What Lady Gaga will wear to the Grammy awards. Maybe God is @LadyGaga's new designer ;)
I have just watched so many videos of Lady Gaga live, she's amazingggggg ;]
Barack Obama's Facebook news feed for the past two weeks. - By ...: Barack Obama was tagged in a photo: http://bit.ly/gsnGV2

Table 6.13: Example of tweets from topic *<justin bieber>*, *<lady gaga>* and *<barack obama>*

Absolute weight often detects short topics that are general and frequently appeared on a daily basis. These topics are mostly greeting messages such as *<happy birthdai>*, where users send happy birthday tweets to the others; personal babble such as *<ic cream>*, where users post tweets when they are eating ice cream. Short topics often occupy the top 10 positions and the tweets contained within these topics are rarely related to news.

Topic	Description	Count
<justin bieber>	Discussion about singer Justin Bieber.	251
<blog post>	Users tweet when they post a new blog.	231
<happi birthdai>	Users sending birthday messages.	211
<check video>	Users posts url about a new video.	128
<hate peopl>	Using complaining about people they hate.	122

Table 6.14: Example of top topics detected using absolute support

Other periodic topic such as *<hate mondai, mondai morn>* that contain tweets such as:

i hate monday mornings.

and similarly, *<happi fridai>*, *<fridai night>*. These topics occur periodically and are mostly about users express their personal feelings.

Relative Weight

Relative weight considers pattern importance in a tweet; therefore, topics detected using relative weight are longer patterns that are considered more important. Relative weights address the problem of absolute support by reducing short generic topics, however, topics detected using relative weights are not immune to noise and spam.

Most long topics contain tweets that are similar. For instance, horoscope topic *<person life merg profession world bu virgo>* and self-promoting topic *<send text photo video media phone free>*; these topics often contain 20 to 40 tweets that are similar, such as:

> *Ping me on @pingchat at ID: dawsooon - Send text, photos, videos, and other media to my phone for free! http://pingchat.com*

Most of the tweets in this topic are similar, except the ID (e.g. *dawsoon*) is different. This type of tweet is not authored by genuine users and is mostly generated automatically by third party apps or websites.

Position	Topics
1	watch werevertumorro video la escuela es para tonto
3	moon current visit hous friend ass aquariu
4	stop think ur missin start thinkin stuff
6	perform job feel exception satisfi cancer
8	post photo facebook album

Table 6.15: Example of topics detected using relative weight ($k \leq 20$)

Other auto-generated tweets can be posted by games or gamification services. These tweets confuse the news detection algorithm to detect topics such as *<reach lvl beat*

game> and <*fav tweet favorit peopl*>. These advertisement tweets try to attract other users in participating games or using services, and contain tweets such as:

> *Hey, I just reached Lvl 3 in #MobsterWorld Beat me in the game! http://www.playmobsterworld.com/?platform=twitter&source=oneline_levelup*

and,

> *@jane_bot, 5 Favs! Your tweet has been favorited by 5 people. http://favstar.fm/t/33676476364619776*

These noisy topics often lead to poor detection performance in top positions (Table 6.15).

Inspecting the top five topics commonly detected using Relative Weight, it shows that most of the tweets are composed by different users. This means that we are unable to eliminate these topics, even if we track the author of each tweet. News related topics are often detected only after top 20 position (Table 6.16).

Position (k)	**Topic**	**Description**
20	<*user egypt bypass twitter facebook block*>	Users in Egypt bypass Twitter and Facebook blocks.
21	<*nomin nobel peac prize*>	Julian Assange nominated for Nobel Peace Prize.
27	<*jimmi buffett fall stage*>	Jimmy Buffett Falls Off Stage During a Concert in Sydney, Australia.
34	<*microsoft sold million kinect*>	Microsoft sells 10 million Kinect devices.
31	<*tear gas canister*>	Controversial issue of making tear gas canisters in the USA.

Table 6.16: Examples of topics detected using relative weights ($k \geq 20$)

Burstiness

Burstiness represents sudden spikes in time series data, which suits the nature of news

detection task. Topics detected using burstiness contain mostly tweets about celebrities and their new movie or new song release, and tweets that critcize public figures such as politicians and sports players.

During the period of NBA all-star balloting, burstiness captured related topics such as basketball players *<rai allen>*, *<jason kidd>* and the player roster issue *<kevin love replac yao ming>*. Burstiness also captured commercial topic *<taco bell>*, when a lawyer firm filed a lawsuit against the fast-food chained restaurant Taco Bell, questioning the beef content in their food products. Another topic is *<richard kei andi grai>*—an incident about the suspension of Sky Sports presenters Richard Keith and Andy Gray after their inappropriate comments during a show. Burstiness measures the change in quantity of tweets and is not affected by the pattern weights, thus addressing the problem of absolute support by lowering the ranking of popular topics with a high volume (Figure 6.8). These topics not detected at top positions using burstiness, which shows that burstiness is capable of suppressing the effect of popular topics that would otherwise be considered as trending topics by using absolute support.

Burstiness misclassifies entertainment-related topics such as *<watch extrem home makeov>*—users tweet while watching TV programme Extreme Home Makeover, and *<watch movi>*—users tweet about the movie they are watching. Burstiness also detects short topic such as *<pink friday>*—a new album released by singer Nicky Minaj, and long topics such as *<beauti god make mistak track babi born>*—users shared the lyrics of a new song by singer Lady Gaga, when she was about to release her new singles "Born This Way" during late January 2011. Bursty topics related to entertainment are difficult to be judged as their news value might be lower compared to news of other genre. Tweets contained in these topics will not be filtered by the pre-processing classifier either, as most of the tweets have appropriate length, a rich set of keywords and relevant hashtags.

Performance of burstiness is also affected by other non-news topic such as advertising

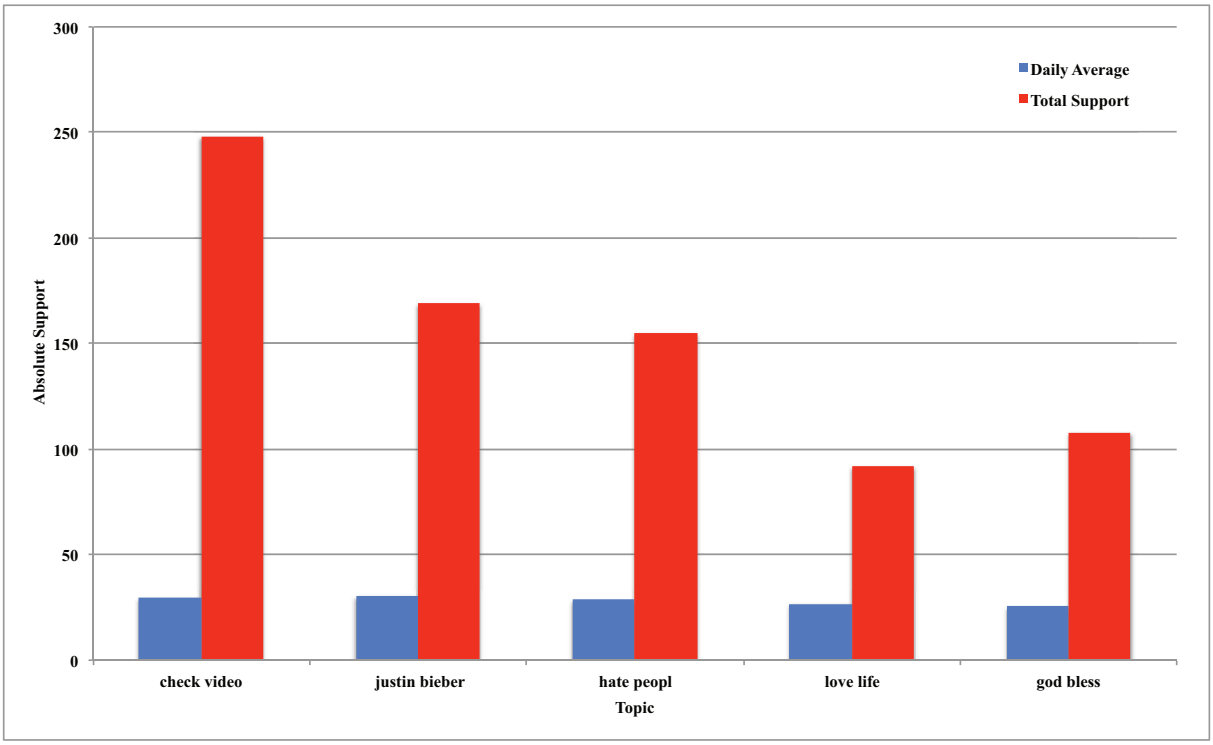


Figure 6.8: Topics with high daily total support compared with daily average of other topics on the same day

and temporal sensitive topics. An instance is the horoscope topic *<overli enthusiast fun leo>*, which contains the tweet:

> *You are overly enthusiastic about having fun today, even if it... More for Leo http://twittascope.com/?sign=5*

This is a popular topic type, as horoscope service websites release tweets of daily predictions for different horoscopes and zodiac signs at particular times of a day. Time-related topics such as *<lion sleep tonight>*, *<goin bed>* and *<cook dinner>* will trigger multiple bursts at different time of the day, depending on the time zone of the users. These topics display a bursty characteristic but are not related to news.

Sentiments

Another key aspect in news is that of public opinion. Sentiments are expected to represent the level of public interest, capturing topics that contain strong sentiments by

measuring opinionated information. However, results show that sentiment analysis alone is insufficient as a feature to capture meaningful news topics.

Topics	Average Tweet Length
<*happi birthdai love*>	4.40
<*ugh hate*>	4.60
<*hate mondai*>	3.97
<*morn beauti*>	4.61
<*happi dai*>	5.42

Table 6.17: Example of top topics detected using sentiment features

Sentiment feature often detects short topics at top positions (Table 6.17). In fact, nearly all the top 50 topics detected using only sentiments score are short and meaningless. For instance, <*morn beauti*>, contains the following tweet:

Good Morning Beautiful People! :)

and,

Morning everyone, make it a beautiful day.

We believe this is caused by the sentiment calculation algorithm which normalizes the sentiment score based on the length, therefore the effect of short tweets are amplified. Detecting news topics that contain sentiment terms seems almost impossible and unlikely to be viable, as news topics do not normally contain opinionated terms within the topic.

Hashtags

Hashtagging is the usual way Twitter users label topics and events. Hashtags associated with events are either assigned officially or formed naturally within the user communities. News topics detected by hashtags include tennis players <*li, na*> and <*andi, murray*>,

who were mentioned in tweets during Australia Tennis Open 2011, with hashtags *#ausopen*; topics related to the Egyptian protest such as <*tear, ga*> and <*secur, forc*> tagged under *#jan25* (as the Egyptian Revolution was from 25th January).

One problem of using the hashtag count is that hashtags are sensitive to noise. Spammers exploit hashtags by including trending hashtags in tweets, for example:

> *Social Dominance: What Newborn Babies Are Teaching Us About #Leadership: http://wp.me/pVReH-Oh #relationships #superbowl #jan25*

This type of spamming tweet often contains multiple popular hashtags to increase its findability, or contains a url that leads to websites containing malware. Also, contents from tweets containing such hashtags are usually irrelevant to the hashtags.

By counting hashtags, we also risk introducing more noise as sometimes users would tag multiple tweets with multiple hashtags. Hashtag itself is insufficient for detecting news topics as its performance are easily affected by noise and requires a reliable approach to ensure content quality.

Urls

Including urls within tweets is the only way to provide additional content beyond the 140 characters limit. News-related topics often contain urls that link to news sources such as CNN, Wall Street Journal, and BBC. Other alternatives include Mashable.com, Techcrunch.com and Endgadget.com, which are digital media with online versions only.

From our experiments, we found that topics detected by counting original urls in tweets without any processing did not produce any sensible results. These topics are noisy, and link to urls that contain irrelevant information. Two particular topics are <*check video*>, which contains the tweet:

> *Check this video out – Dwele - Find A Way http://t.co/wUS5mpp via @youtube*

and *<blog post>* that contains the tweet:

New blog post: How to Select a Good Wine http://blog.studentsagain.com/?p=9644

These tweets are generated from users who install a publish plugin[2] that sends tweets automatically when a new blog post is composed or a new video is uploaded.

Url	Tweet
http://www.4sq.com	I'm at Wyse Technology (3471 North First St, San Jose) http://4sq.com/egxLrH
http://www.blogspot.com	Check out my new blog http://fashionfyoozdculture.blogspot.com Suggestions would be valued
http://www.facebook.com	I posted 8 photos on Facebook in the album "New Paintings" http://fb.me/zeIf6XsB
http://www.wordpress.com	Special Monday morning thanks to all of our new followers! Visit http://pgoc.wordpress.com for more information on our upcoming events
http://www.linkedin.com	InMaps - I visualized my LinkedIn network http://t.co/4fnY5wU

Table 6.18: Example of tweet with social media url

Many social media applications provide similar functionality when users perform any activities such as posting comments, checking in at new location and uploading new media. This has caused topics such as *<post photo facebook album>* to be mistaken as a news topic given its high url count.

Another type of noisy topic detected by using url comes from information and marketing (Table 6.19). These topics contain tweets that link to the same url: dedicated machines (or "bots") are set up to regularly publish urls about information such as weather forecast, or product promotions.

[2]http://wordpress.org/plugins/twitter-auto-publish/

Topic	Tweet Example
<*vote photo*>	I just voted for this photo http://plixi.com/p/70486300
<*airport tx wind*>	Fog/Mist and 52 F at Orange County Airport, TX Winds are Calm. The humidity is 100%.,Last Updated on Jan 31 2011, 4:45 http://s2z.us/mk.htm
<*weight loss*>	The 7 Day Diet Plan for weight loss http://dld.bz/kX8S
<*free ship*>	Technical Pro Podcaster Kit for $64 + free shipping http://bit.ly/g4nDLB
<*enter win*>	The Frugal Ya-Ya: Enter to win tickets to the Super Bowl! http://t.co/iLotT2V

Table 6.19: Example of topics related to marketing and weather forecasts

To address this issue, we use a metric to measure the spread of urls within a topic. Url spread measures how many unique urls are shared within a topic, and reduce the significance of topics that mostly contain many identical urls, which are likely to be spam topics. Due to the word limit, urls in tweets are often compressed by url shortening services such as *http://bit.ly* and *http://t.co*, therefore the urls need to be decompressed into their original form before measuring. This is to prevent a url that is compressed by different shortening services from appearing as multiple unique urls.

One way to interpret topics detected by url spread is that these topics are interesting, as users are actively sharing and contributing information from multiple sources. The cases where url spread is unable to perform are when media outlets are considered as topics, for instance <*ny time*> and <*wall street journal*>. These topics often contain tweets that link to different articles of the publishers, for example <*new york time*>:

NY Times with a brief article on Clowney and Clemson - http://nyti.ms/ff4Xli

and Wall Street Journal:

> *Ripley SA : Chile Stocks Rising 1.1% Early, Following January Slump - Wall Street Journal http://uxp.in/27682206*

These topics contain many unique urls which are often misclassified as topics using url spread. Either manual inspection or content similarity measurement is required to further evaluate news relevance for these topics.

Retweet

A retweet is a unique Twitter action where users forward a copy of tweet from another user to their own followers. Retweets imply user recognition with the tweet content, finding it valuable and therefore sharing it with their own followers. In most cases, news topics are only detected within the top 20 positions, as retweets for topics beyond these drop significantly after. Retweet counts are low for most topics, with each topic containing only one to two retweets on average.

Retweets are able to capture meaningful topics for international topics and critical issues. For instance, during the Egyptian protest, users actively retweeted information about different aspects of the riot, such as:

> *RT @octavianasr: Hearing UNCONFIRMED info that the Egyptian Army getting ready to announce President Hosni Mubarak is stepping down.*

Users are also motivated to retweet controversial issues such as *<health care law unconstitut>* with tweet:

> *RT @foxnewspolitics: Florida judge rules that health care law is unconstitutional, says entire act must be declared void*

and emergency information such as *<due sever weather>* with tweet:

> *RT @msichicago: Due to severe weather, MSI closing @ 2 PM and will be closed all day Wed 2/2. Stay tuned for updates; everybody stay safe.*

Topics that are related to announcements can also be captured by retweets: *<jai carnei>*—tweets related to Jay Carney, who is selected to be White House secretary.

Topics detected using retweets are mainly entertainment topics, where users retweet the release of new albums, movies and news about celebrities, for instance:

> *RT @LadyGaGaWatch: French Montana Drops Video, Looks To Record With Lady Gaga: http://bit.ly/fsFraW*

This sort of news, although contributing to news topics, has much less importance and impacts, from a journalistic perspective.

Retweeting is also actively employed by companines to promote their products through marketing campaigns that encourage users to retweet in order to be eligible to enter a lucky draw:

> *RT @Gwen_UsBeauty NEW GAME! Play BEAUTY ROULETTE! Follow and RT to enter to win something random yet covetable from the Beauty Closet! Go!!!*

Another problem that using retweet is facing is that many tweets are in fact examples where users are self-publicizing, seeking for followers, and they will post a tweet such as:

> *Hey guys please RT this So I can get some more followers thanks love you all xxx.*

This type of tweet is not related to news and the "RT" does not represent a genuine retweet, often confusing the system.

6.5.4 Results of News Detection using PMM (NDPMM)

The results of News Detection using Pattern Model for Microblogs (NDPMM) highlight one of the key contributions of this thesis. NDPMM presents a single computational model that combines multiple features to compute a relevance score for each topic. NDPMM uses a logistic regression model, trained using a set of 2000 manually evaluated topics to

learn the coefficients of relative weight, burstiness, sentiments, hashtag, url and retweet.

The results show that NDPMM performs better than using any single feature alone. Combining multiple features captures news topics related to international incidents, sports, politics and entertainment (Table 6.20). Along with these popular topics, it also detects news topics related to technology such as *<microsoft sold million kinect>*—sales of Microsoft Kinects during holiday, *<kindl book outsel paperback>*—Kindle books sell more than paperback books, and *<android honeycomb sdk>*—the release of Android Honeycomb SDK. Technology-related news is mainly reported by digital media such as Mashable.com, Engadget.com and CNET.com.

Topic	**Description**	**Relevance Score**
<tahrir sq>	Tahrir square, the main protest destination during 2011 Egyptian revolution	2.524
<torr chelsea deal>	Liverpool football club's striker Fernando Torres transferred to Chelsea Football Club	1.227
<olbermann announc departur msnbc>	MSNBC Host Keith Olbermann announced his resignation	0.518
<chicago car salesman fire green bai packer tie>	A Chiago based car salesman was fired for wearing tie of the opposition team his boss supported.	0.445
<start american forbidden serv countri love>	A popular quote from Barack Obama's State of The Union speech	0.198

Table 6.20: Example of news topics detected using multiple features

The motivation behind multiple feature combination is to improve overall detection performance by taking advantage of the benefits of individual features. Results show that combining multiple features eliminates high volume topics such as *<lady gaga>*, *<justin bieber>*; short and bursty topics *<good night>*, *<happi dai>*; and noisy topics with tweets that contain duplicate urls such as *<check video>* and *<weight loss diet plan>*.

Although most of the irrelevant topics are filtered, there are still topics that are difficult to be judged. For instance, location topics *<hong kong>* and *<unit kingdom>* are detected as news relevant topics. These topics contain tweets with news information; however, the focus of the topics are not clear and therefore should not be considered as news topics (Table 6.21).

Tweets
(UDN) Immigration officer fired after putting wife on list of terrorists to stop her flying home: UNITED KINGDOM... http://bit.ly/hq6DBr
Hong Kong Herald: China, charmer and bully: While admitting that there were differences of opinion on... http://bit.ly/f1kNFl #China #HK
The Difference between the United Kingdom, Great Britain, and England explained. http://youtu.be/rNu8XDBSn10
#jobs #careers #London Automated Test Lead Leading Retail Ecommerce .Com Qtp Qc (United Kingdom,City of London) http://bit.ly/haPwPv
NYTimes article on HK's 3+3+4 transition: Hong Kong's Universities Decide Bigger Is Better - http://nyti.ms/gPXn6t #westhk #edchat

Table 6.21: Example of tweets from topic *<unit kingdom>* and *<hong kong>*

On the other hand, another location topic *<moscow domodedovo airport>*, which is about a suicidal bombing happened on 24th January 2011, is considered a news topic as the topics are more specific and the tweets have more focus (Table 6.22).

Other misclassified topics include *<blog post>*, where most tweets in this topic contain unique urls from various blogging platforms, and there is not enough significance from other features to help determining news relevance; for *<free ship>* and *<weight loss>*, these topics contain lengthy tweets, unique urls, hashtags and retweets and are therefore considered as news-related topics. The news detection algorithm is unable to handle these topics accurately, based on the combination of multiple features only, so they may require higher level processing for news relevance assessment.

Tweets
Russian media reporting suicide bomber carried out attack at arrivals hall of Moscow's Domodedovo Airport that killed at least 20
Blast at Moscow's Domodedovo airport was suicide attack - source citing initial probe #news
MT @BBCBreaking: Russian media reporting a suicide bomber killed at least 10 people at Moscow's Domodedovo Airport
Russian media now reporting that at least 31 people were killed and 130 injured in bombing at Moscow's Domodedovo airport, from AFP via BBC
Moscow's Domodedovo airport - the busiest in the Russian capital - is hit by e-mailxplosion with 10 ppl reported killed. - BBC website

Table 6.22: Example of tweets from topic *<moscow domodedovo airport>*

6.6 Summary

From our experiments, we found that representing tweets using sequential patterns improves the performance of both retrieval and news topic detection tasks. These tasks are challenging as we are dealing with large amounts of fragmentized, noisy and unstructured data. This requires a huge effort to process, clean and extract features before any useful data can be mined.

Noise in tweets and their short length are the main problems in microblog processing tasks. For retrieval, by using pattern representation, the relationship between terms is captured to help to disambiguate the meaning of terms in the query. We further show that by using query expansion combined with Twitter specific features, the retrieval performance is significantly improved.

For news topic detection, the primary aim is to find potential news topics from tweets. Results show that the multiple features combination performs better than using any single feature only. Relative weight helps to remove short patterns that do not contribute much towards topics; burstiness captures topics that show a sudden increase "spike", and

eliminate topics that are overly popular; sentiments, while not useful when used as a single feature, help to identify topics that draw public interest; Twitter features model different types of user activities while spreading news related information.

The news topic detection algorithm is able to eliminate meaningless and noisy topics in most cases. The nature of the social media data has caused the detected topics to be mostly related to entertainment, politics and technology. The topics are also affected by user demographics. We believe that the algorithm can be further improved by implementing other rules to eliminate topics that are irrelevant. The algorithm can also be implemented into other detection and monitoring applications by adjusting the parameters according to the requirements of each application.

Chapter 7

Conclusion and Future Work

It is evident that large amounts of news topics can be detected from microblogs. Many related studies have been conducted on finding trending topics and events, but there are still few studies for generic news. This thesis presents a news topic detection framework to address this problem and to contribute to the text mining field using microblog data collected from Twitter.

Tweet is a document type with strong unique characteristics. Its short nature has made it into a topic-focused document. When posting a news-related tweet, the topic shows a close correlation with temporal information. Users often express personal opinions and sentiments while using tweets to participate in discussion.

This thesis presents a novel news topics detection framework, based on sequential pattern mining and consider Twitter characteristics. The framework has three components: feature extraction, topic detection and news topic identification. For features, tweets are represented using patterns to overcome the limitations of terms. The vast majority of previous studies on microblogs built on the success of term-based techniques, but they do not consider the relationship between terms, which therefore is more sensitive to noise and which leads to ambiguity. This has caused query mismatch that affects retrieval and detection performance.

Sequential patterns capture meaningful word groups that play a significant role in news topics. Such patterns represent word groups in proper order and reduce ambiguity using sequential information. Through experiments, we showed that patterns achieve better retrieval results than term-based representations. Furthermore, sequential patterns outperform non-sequential patterns of different weighting methods.

Two main problems encountered while mining patterns from tweets are the large amount of redundant patterns generated and the low support problem for long patterns. Redundant patterns are inevitable in any pattern mining process, but the noise in tweets worsens the problem. To overcome this problem, we consider the part-of-speech tags to eliminate patterns that do not contribute much to topic understanding. We also observe that most patterns from short tweets are less informative. By considering these factors, the algorithm captures more readable and representative topics for tweets.

Another pertinent issue addressed is that useful long patterns with high specificity suffers from low support. In contrast, patterns with high support are mostly short patterns that do not contain much information. We present an algorithm that computes pattern weights in a tweet according to their importance by measuring the amount of information they contain. This calculation significantly reduce the effect of short noisy patterns, and improves the performance of news detection.

Sequential patterns are also the basis of our topic detection model. We can consider patterns in tweets as part of a topic, since most tweets are about a single topic. We adopt the text mining techniques used in many knowledge discovery tasks and improve them to fit microblog characteristics.

When being used for information dissemination, Twitter provides unique actions that faciliate the spread of information. These activities improve findability and are used in many critical situations to spread real-time news updates when access to other media is restricted. At the same time they also affect the performance of traditional term-based

processing techniques. Previous studies have used different subsets of features derived from these activities and have achieved promising results by applying them in different topic categories.

On news topic identification, our model combines these Twitter features with other temporal, sentiment and content information, and builds a classification model that evaluates topic relevance score using logistic regression. This score represents the possbility of a topic being related to news. Experimental results show that a multiple feature combination leads to better performance than using any feature alone. Combining features incorporates the benefits of individual features. News detection works well in a large dataset and detects many key news topics during the period of 23rd January 2011 to 9th February 2011.

This presents a framework that is able to detect news from a text medium that is short, unstructured and noisy. Although Twitter is used as the primary testing platform, there seems to be no reason why such framework cannot be applied to other similar short-text content in other social media platforms such as Facebook, Plurk and Tumblr. The generic techniques presented in the framework should be easily tuned for use in other systems to detect different types of topics.

Appendix A

TREC 2011 Microblog Topics

A.1 TREC 2011 Microblog Tracks Topics

Topic	Title	Query Time	Query Tweet Time
MB001	BBC World Service staff cuts	08/02/2011 12:30	34952194402811900
MB002	2022 FIFA soccer	08/02/2011 18:51	35048150574039000
MB003	Haiti Aristide return	08/02/2011 21:32	35088534306033600
MB004	Mexico drug war	02/02/2011 17:22	32851298193768400
MB005	NIST computer security	04/02/2011 17:44	33581589627666400
MB006	NSA	08/02/2011 16:00	35005178885181400
MB007	Pakistan diplomat arrest murder	8/02/2011 22:56	35109758973255600
MB008	phone hacking British politicians	7/02/2011 17:42	34668458591395800
MB009	Toyota Recall	8/02/2011 21:41	35090855064764400
MB010	Egyptian protesters attack museum	29/01/2011 20:06	31443107291598800
MB011	Kubica crash	6/02/2011 10:38	34199299428581300
MB012	Assange Nobel peace nomination	31/01/2011 21:02	32181966761631700
MB013	Oprah Winfrey half sister	24/01/2011 15:43	29565006546735100
MB014	release of The Rite	2/02/2011 12:31	32778015167479800
MB015	Thorpe return in 2012 Olympics	30/01/2011 12:20	31688182005235700

Table A.1: TREC 2011 Microblog Dataset Topics MB001 - MB025

Table A.2: TREC 2011 Microblog Dataset Topics MB026 - MB050

Topic	Title	Query Time	Query Tweet Time
MB016	release of Known and Unknown	24/01/2011 17:03	29585186899365800
MB017	White Stripes breakup	2/02/2011 19:13	32879343399084000
MB018	William and Kate fax save-the-date	26/01/2011 8:59	30188073790742500
MB019	Cuomo budget cuts	7/02/2011 23:25	34754540519563200
MB020	Taco Bell filling lawsuit	6/02/2011 7:09	34146608102772700
MB021	Emanuel residency court rulings	29/01/2011 3:03	31185639047172000
MB022	healthcare law unconstitutional	1/02/2011 22:17	32563233118224300
MB023	Amtrak train service	8/02/2011 20:04	35066441501900800
MB024	Super Bowl seats	8/02/2011 17:11	35022813232373700
MB025	TSA airport screening	3/02/2011 19:52	33251413001764800
MB026	US unemployment	4/02/2011 14:10	33527910379814900
MB027	reduce energy consumption	4/02/2011 4:19	33379210437337000
MB028	Detroit Auto Show	26/01/2011 22:46	30396111764066300
MB029	global warming and weather	8/02/2011 1:05	34779934836785100
MB030	Keith Olbermann new job	8/02/2011 22:51	35108366829232100
MB031	Special Olympics athletes	4/02/2011 8:44	33445664922800100
MB032	State of the Union and jobs	4/02/2011 2:08	33346093525762000
MB033	Dog Whisperer Cesar Millan techniques	27/01/2011 19:27	30708594202648500
MB034	MSNBC Rachel Maddow	4/02/2011 22:42	33656631187210200
MB035	Sargent Shriver tributes	24/01/2011 7:18	29437816727404500
MB036	Moscow airport bombing	24/01/2011 23:00	29674954899333100
MB037	Giffords recovery	3/02/2011 18:05	33224462191038400
MB038	protests in Jordan	1/02/2011 12:46	32419560749531100
MB039	Egyptian curfew	28/01/2011 18:14	31052423128686500
MB040	Beck attacks Piven	31/01/2011 20:33	32174687102435300
MB041	Obama birth certificate	31/01/2011 17:55	32134993337647100
MB042	Holland Iran envoy recall	7/02/2011 20:47	34714824982134700
MB043	Kucinich olive pit lawsuit	29/01/2011 8:06	31261786745339900
MB044	White House spokesman replaced	28/01/2011 13:35	30982361281728500
MB045	political campaigns and social media	1/02/2011 12:52	32421023961841600
MB046	Bottega Veneta	8/02/2011 22:34	35104330025541600
MB047	organic farming requirements	8/02/2011 0:12	34766556445540300
MB048	Egyptian evacuation	31/01/2011 9:36	32009428471386100
MB049	carbon monoxide law	1/02/2011 22:44	32569981321347000
MB050	war prisoners Hatch Act	25/01/2011 2:13	29723425576587200

Appendix B

Full Evaluation Results

Please see the next page.

Table B.1: Precision @ 30

		Term Baseline		Pattern Baseline					PBMM			
Topic	**TREC Median**	**Lucene**	**TF-IDF**	**NSCP**	**NSCP-PWT**	**NSCP-TWP**	**NSCP-C**	**SCP(C)**	**SCP(J)**	**SCP + Term**	**SCP + T + QE**	**SCP + T + QE + TT**
MB001	0.1409	0.0000	0.6333	0.4333	0.4667	0.4333	0.4333	0.3139	0.6667	0.5333	0.7667	0.9500
MB002	0.0566	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0339
MB003	0.2766	0.0667	0.5667	0.3333	0.2333	0.3667	0.2333	1.0000	0.7000	0.5600	0.7000	0.7283
MB004	0.0453	0.0000	0.1333	0.2667	0.2333	0.4000	0.4000	0.4138	0.4300	0.3467	0.4000	0.9500
MB005	0.4089	0.0667	0.2000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.6045
MB006	0.0902	0.1000	0.2000	0.0000	0.0333	0.0333	0.2333	0.0000	0.0000	0.0000	0.0833	0.5542
MB007	0.1271	0.1000	0.6000	0.2000	0.1000	0.5000	0.5000	0.0000	0.3700	0.2933	0.7667	0.9500
MB008	0.0449	0.0667	0.0000	0.4000	0.5667	0.4000	0.4000	0.5546	0.7000	0.5600	0.7667	0.9500
MB009	0.1386	0.2667	0.6333	0.4667	0.7333	0.7333	0.7333	0.8090	0.7700	0.6133	0.8000	0.9500
MB010	0.0395	0.1667	0.0667	0.0000	0.1333	0.1667	0.1667	0.0081	0.0000	0.0000	0.0000	0.0000
MB011	0.4028	0.1333	0.0333	0.0000	0.0000	0.0667	0.0000	0.0000	0.1700	0.1333	0.0000	0.1583
MB012	0.3385	0.1000	0.0667	0.0333	0.0333	0.0333	0.0333	0.1111	0.7500	0.6000	0.2500	0.4750
MB013	0.2404	0.3333	0.3333	0.0667	0.0667	0.3000	0.1000	0.3333	0.4500	0.3556	0.4074	0.5278
MB014	0.0130	0.1000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.4300	0.3467	0.8667	0.4117
MB015	0.0000	0.0333	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MB016	0.5000	0.0333	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MB017	0.1215	0.3333	0.3333	0.0333	0.0000	0.3333	0.3333	0.3089	0.3000	0.2400	0.4333	0.5067
MB018	1.0000	0.0333	0.0333	0.0333	0.0000	0.0000	0.0000	0.0000	1.0000	0.8000	1.0000	0.9500
MB019	0.1727	0.1333	0.5667	0.2333	0.2333	0.2333	0.2333	0.0958	0.1300	0.1067	0.1667	0.1583
MB020	0.1004	0.0000	0.2333	0.6667	0.5000	0.3667	0.3667	0.1516	0.8000	0.6400	0.8667	0.9500
MB021	0.0870	0.0333	0.4333	0.3000	0.4000	0.3667	0.3667	0.3088	0.4700	0.3467	0.3333	0.7917
MB022	0.1150	0.1333	0.3667	0.4333	0.3667	0.4333	0.4333	0.4308	0.2700	0.2133	0.4333	0.9500
MB023	0.0786	0.0667	0.4000	0.1667	0.2000	0.2000	0.2000	0.2727	0.2300	0.1867	0.4000	0.2533
MB024	0.1188	0.1000	0.0000	0.0000	0.5000	0.5000	0.5000	0.0054	0.2300	0.1867	0.3000	0.9500
MB025	0.0507	0.1000	0.1000	0.0667	0.0333	0.0667	0.0667	0.1000	0.1300	0.1067	0.0333	0.2533
MB026	0.0456	0.2000	0.0667	0.0333	0.3000	0.2000	0.3333	0.0000	0.2300	0.1867	0.0000	0.4433
MB027	0.0208	0.0000	0.1000	0.1000	0.2333	0.1667	0.1667	0.2143	0.1300	0.1067	0.1000	0.1267
MB028	0.1234	0.0667	0.1333	0.0667	0.1333	0.1333	0.1667	0.6667	0.4200	0.3333	0.3333	0.4750
MB029	0.0470	0.1333	0.0000	0.2333	0.2333	0.2333	0.2333	0.1617	0.2000	0.1600	0.3000	0.2533
MB030	0.1533	0.3667	0.2667	0.0000	0.0000	0.3000	0.0000	0.1887	0.0000	0.1600	0.0333	0.2217
MB031	0.1971	0.0667	0.1667	0.0667	0.1000	0.1333	0.1000	0.1667	0.1000	0.5600	0.7000	0.7600
MB032	0.0237	0.1333	0.0000	0.0333	0.1333	0.1667	0.1667	0.0000	0.0000	0.0000	0.0000	0.1583
MB033	0.0038	0.0000	0.0333	0.0333	0.0000	0.0333	0.0333	0.1818	0.0000	0.2000	0.0000	0.2375
MB034	0.0470	0.0667	0.3333	0.0667	0.3333	0.3333	0.3333	0.3333	0.0300	0.3467	0.4333	0.5067
MB035	0.2364	0.2667	0.2333	0.1000	0.0667	0.0333	0.1000	0.6000	0.0000	0.6545	0.6364	0.6909
MB036	0.1295	0.3000	0.5333	0.2333	0.3333	0.3333	0.3333	0.6222	0.3700	0.4533	0.4667	0.9500
MB037	0.1147	0.0333	0.0000	0.0000	0.0000	0.4667	0.4667	0.7500	0.0000	0.5067	0.7000	0.9500
MB038	0.0540	0.0333	0.0333	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5231	0.5769	0.6212
MB039	0.1324	0.1667	0.1667	0.1000	0.0667	0.0667	0.0667	0.0000	0.0000	0.1600	0.0000	0.1267
MB040	0.1546	0.0333	0.1667	0.2000	0.2000	0.2333	0.2333	0.0000	0.0000	0.4000	0.7222	0.4750
MB041	0.0832	0.0333	0.3000	0.3000	0.3333	0.3333	0.3333	0.2133	0.1000	0.3200	0.4333	0.7283
MB042	0.0338	0.0333	0.0333	0.0333	0.1667	0.0333	0.0333	0.0000	0.0000	0.0533	0.0000	0.0633
MB043	0.3242	0.1000	0.6000	0.4333	0.3667	0.3667	0.3667	0.5676	0.0000	0.4533	0.5333	0.7600
MB044	0.0833	0.0667	0.0333	0.2667	0.2333	0.2333	0.2333	0.0538	0.0000	0.1000	0.0000	0.0792
MB045	0.0127	0.1000	0.1333	0.0000	0.0333	0.0667	0.0667	0.0270	0.0000	0.1333	0.1000	0.3167
MB046	0.1380	0.2333	0.2667	0.0667	0.0667	0.1000	0.1000	0.4211	0.0000	0.4923	0.6154	0.5846
MB047	0.0051	0.0000	0.0000	0.0667	0.1000	0.0333	0.0333	0.0000	0.0000	0.0000	0.0714	0.2036
MB048	0.0301	0.0000	0.0667	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
MB049	0.0625	0.0000	0.0333	0.0333	0.0333	0.0333	0.0333	0.0294	0.0000	0.4000	0.5000	0.4750
MAP	**0.1421**	**0.1007**	**0.1966**	**0.1347**	**0.1694**	**0.2034**	**0.1973**	**0.2330**	**0.2159**	**0.2831**	**0.3475**	**0.4942**

Keys: C - Cosine Similarity, J - Jaccard Index, T - Terms Only, TT - Twitter Features

Appendix C

Example of Twitter API output

C.1 Twitter API JSON Output


```
"coordinates": null,
"created_at": "Sat Sep 10 22:23:38 +0000 2011",
"truncated": false,
"favorited": false,
"id_str": "112652479837110273",
"entities": {⊞ ⋯},
"in_reply_to_user_id_str": "783214",
"text": "@twitter meets @seepicturely at #tcdisrupt cc.@boscomonkey @episod http://t.co/6J2EgYM",
"contributors": null,
"id": 112652479837110270,
"retweet_count": 0,
"in_reply_to_status_id_str": null,
"geo": null,
"retweeted": false,
"possibly_sensitive": false,
"in_reply_to_user_id": 783214,
"place": null,
"source": "<a href=\"http://instagr.am\" rel=\"nofollow\">Instagram</a>",
"user": {⊞ ⋯},
"in_reply_to_screen_name": "twitter",
"in_reply_to_status_id": null
```


Figure C.1: Twitter Specific Entities


```
"user": {
    "profile_sidebar_border_color": "eeeeee",
    "profile_background_tile": true,
    "profile_sidebar_fill_color": "efefef",
    "name": "Eoin McMillan ",
    "profile_image_url": "http://a1.twimg.com/profile_images/1380912173/Screen_shot_2011-06-03_at_7.35.36_PM_normal.png",
    "created_at": "Mon May 16 20:07:59 +0000 2011",
    "location": "Twitter",
    "profile_link_color": "009999",
    "follow_request_sent": null,
    "is_translator": false,
    "id_str": "299862462",
    "favourites_count": 0,
    "default_profile": false,
    "url": "http://www.eoin.me",
    "contributors_enabled": false,
    "id": 299862462,
    "utc_offset": null,
    "profile_image_url_https": "https://si0.twimg.com/profile_images/1380912173/Screen_shot_2011-06-03_at_7.35.36_PM_normal.png",
    "profile_use_background_image": true,
    "listed_count": 0,
    "followers_count": 9,
    "lang": "en",
    "profile_text_color": "333333",
    "protected": false,
    "profile_background_image_url_https": "https://si0.twimg.com/images/themes/theme14/bg.gif",
    "description": "Eoin's photography account. See @mceoin for tweets.",
    "geo_enabled": false,
    "verified": false,
    "profile_background_color": "131516",
    "time_zone": null,
    "notifications": null,
    "statuses_count": 255,
    "friends_count": 0,
    "default_profile_image": false,
    "profile_background_image_url": "http://a1.twimg.com/images/themes/theme14/bg.gif",
    "screen_name": "imeoin",
    "following": null,
    "show_all_inline_media": false
},
```


Figure C.2: Twitter User Details

Literature Cited

Abel, F., Gao, Q., Houben, G.-J., and Tao, K. (2011). Analyzing user modeling on twitter for personalized news recommendations. In Konstan, J., Conejo, R., Marzo, J., and Oliver, N., editors, *User Modeling, Adaption and Personalization*, volume 6787 of *Lecture Notes in Computer Science*, pages 1–12. Springer Berlin Heidelberg.

Abel, F., Hauff, C., Houben, G.-J., Stronkman, R., and Tao, K. (2012). Twitcident: fighting fire with information from social web streams. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 305–308, New York, NY, USA. ACM.

Agarwal, P., Vaithiyanathan, R., Sharma, S., and Shroff, G. (2012). Catching the long-tail: Extracting local news events from twitter. In Breslin, J. G., Ellison, N. B., Shanahan, J. G., and Tufekci, Z., editors, *ICWSM*. The AAAI Press.

Ahmad, A. N. (2010). Is twitter a useful tool for journalists? *Journal of Media Practice*, 11(2):145–155.

Ahmed, A. (2011). Thousands protest in egypt. http://edition.cnn.com/2011/WORLD/meast/01/25/egypt.protests. Last accessed on Nov 01, 2013.

Albakour, M.-D., Macdonald, C., and Ounis, I. (2013). Identifying local events by using microblogs as social sensors. In *Proceedings of the 10th Conference on Open Research Areas in Information Retrieval*, OAIR '13, pages 173–180, Paris, France, France. Le Centre de Hautes Études Internationales d'Informatique Documentaire.

Algarni, A., Li, Y., Xu, Y., and Lau, R. Y. (2009). An effective model of using negative relevance feedback for information filtering. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1605–1608, New York, NY, USA. ACM.

Allan, J., Harding, S., Fisher, D., Bolivar, A., Guzman-Lara, S., and Amstutz, P. (2005). Taking topic detection from evaluation to practice. In *System Sciences, 2005. HICSS '05. Proceedings of the 38th Annual Hawaii International Conference on*, pages 101a–101a.

Allan, J., Papka, R., and Lavrenko, V. (1998). On-line new event detection and tracking. In *Proceedings of the 21st Annual International ACM SIGIR Conference*

on Research and Development in Information Retrieval, SIGIR '98, pages 37–45, New York, NY, USA. ACM.

Allen, N. (2011). Charlie sheen rushed to hospital 'after party'. http://www.telegraph.co.uk/news/worldnews/northamerica/usa/8287550/Charlie-Sheen-rushed-to-hospital-after-party.html. Last accessed on Nov 01, 2013.

Alonso, O., Carson, C., Gerster, D., Ji, X., and Nabar, S. U. (2010). Detecting Uninteresting Content in Text Streams. In Lease, M., Carvalho, V., and Yilmaz, E., editors, *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)*, pages 39–42, Geneva, Switzerland.

Asur, S. and Huberman, B. (2010). Predicting the future with social media. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 492–499. IEEE.

Barbosa, L. and Feng, J. (2010). Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44, Stroudsburg, PA, USA. Association for Computational Linguistics.

Becker, H., Naaman, M., and Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Benhardus, J. and Kalita, J. (2013). Streaming trend detection in twitter. *IJWBC*, 9(1):122–139.

Bermingham, A. and Smeaton, A. F. (2010). Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1833–1836, New York, NY, USA. ACM.

Bernstein, M. S., Suh, B., Hong, L., Chen, J., Kairam, S., and Chi, E. H. (2010). Eddi: interactive topic-based browsing of social status streams. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 303–312, New York, NY, USA. ACM.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.

Bollen, J., Pepe, A., and Mao, H. (2009). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583.

Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 2010 43rd Hawaii International Conference on System Sciences*, HICSS '10, pages 1–10, Washington, DC, USA. IEEE Computer Society.

Brody, S. and Diakopoulos, N. (2011). Cooooooooooooooollllllllllllll!!!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 562–570, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bruns, A. and Burgess, J. (2011). # ausvotes: How twitter covered the 2010 australian federal election. *Communication, Politics and Culture*, 44(2):37–56.

Bruns, A., Burgess, J. E., Crawford, K., and Shaw, F. (2012). # qldfloods and@qpsmedia: Crisis communication on twitter in the 2011 south east queensland floods.

Bruns, A. and Highfield, T. (2012). Blogs, twitter, and breaking news: the produsage of citizen journalism. *Produsing Theory in a Digital World: The Intersection of Audiences and Production in Contemporary Theory*, 80:15–32.

Bulearca, M. and Bulearca, S. (2010). Twitter: a viable marketing tool for smes. *Global Business and Management Research: An International Journal*, 2(4):296–309.

Burns, A. and Eltham, B. (2009). Twitter free iran: an evaluation of twitter's role in public diplomacy and information operations in iran's 2009 election crisis. In *Record of the Communications Policy & Research Forum*, pages 298–310.

Castillo, C., Mendoza, M., and Poblete, B. (2011). Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 675–684, New York, NY, USA. ACM.

Cataldi, M., Di Caro, L., and Schifanella, C. (2010). Emerging topic detection on twitter based on temporal and social terms evaluation. In *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, MDMKDD '10, pages 4:1–4:10, New York, NY, USA. ACM.

Chang, H.-C. (2010). A new perspective on twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4.

Cheong, M. and Lee, V. (2009). Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base. In *SWSM '09: Proceeding of the 2nd ACM workshop on Social Web Search and Mining*, pages 1–8, New York, NY, USA. ACM.

Cheong, M. and Lee, V. (2011). A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. *Information Systems Frontiers*, 13(1):45–59.

CNN (2011). Medvedev boosts security after airport terror attack. http://edition.cnn.com/2011/WORLD/europe/01/24/russia.airport.explosion/. Last accessed on Nov 01, 2013.

Correa, D. and Sureka, A. (2011). Mining tweets for tag recommendation on social media. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 69–76, New York, NY, USA. ACM.

Cui, A., Zhang, M., Liu, Y., and Ma, S. (2011). Are the urls really popular in microblog messages? In *Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on*, pages 1–5.

Cui, A., Zhang, M., Liu, Y., Ma, S., and Zhang, K. (2012). Discover breaking events with popular hashtags in twitter. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 1794–1798, New York, NY, USA. ACM.

Culotta, A. (2010a). Detecting influenza outbreaks by analyzing twitter messages.

Culotta, A. (2010b). Towards detecting influenza epidemics by analyzing twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 241–249, Stroudsburg, PA, USA. Association for Computational Linguistics.

Deuze, M. (2005). What is journalism?: Professional identity and ideology of journalists reconsidered. *Journalism*, 6(4):442–464.

Deuze, M. and Marjoribanks, T. (2009). Newswork. *Journalism*, 10(5):555–561.

Diakopoulos, N. and Shamma, D. (2010). Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM.

Diao, Q., Jiang, J., Zhu, F., and Lim, E.-P. (2012). Finding bursty topics from microblogs. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 536–544, Stroudsburg, PA, USA. Association for Computational Linguistics.

Du, Y., He, Y., Tian, Y., Chen, Q., and Lin, L. (2011). Microblog bursty topic detection based on user relationship. In *Information Technology and Artificial Intelligence Conference (ITAIC), 2011 6th IEEE Joint International*, volume 1, pages 260–263.

Editor CNN (2011). Illinois supreme court keeps emanuel on ballot. http://edition.cnn.com/2011/POLITICS/01/27/emanuel.ballot. Last accessed: 01 Nov, 2013.

Efron, M. (2010). Hashtag retrieval in a microblogging environment. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 787–788, New York, NY, USA. ACM.

Efron, M. (2011). Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008.

Efron, M. and Golovchinsky, G. (2011). Estimation methods for ranking recent information. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 495–504, New York, NY, USA. ACM.

Efron, M., Organisciak, P., and Fenlon, K. (2012). Improving retrieval of short texts through document expansion. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '12, pages 911–920, New York, NY, USA. ACM.

Farhi, P. (2009). The twitter explosion-whether they are reporting about it, finding sources on it or urging viewers, listeners and readers to follow them on it, journalists just can't seem to get enough of the social networking site. just how effective is it as a journalism tool? *American Journalism Review (AJR)*, 31(3):26.

Feldman, R. and Sanger, J. (2006). *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, New York, NY, USA.

Gabbatt, A. (2011). Oprah winfrey reveals 'miracle' half-sister. http://www.theguardian.com/tv-and-radio/2011/jan/25/oprah-winfrey-miracle-half-sister. Last accessed on Nov 01, 2013.

Gaffney, D. (2010). #iranelection: Quantifying online activism. In *Web Science Conference*, Raleigh, NC, USA.

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., and Kellerer, W. (2010). Outtweeting the twitterers - predicting information cascades in microblogs. In *Proceedings of the 3rd conference on Online social networks*, WOSN'10, pages 3–3, Berkeley, CA, USA. USENIX Association.

Gamon, M., Basu, S., Belenko, D., Fisher, D., Hurst, M., and König, A. C. (2009). Blews: Using blogs to provide context for news articles. In Adar, E., Hurst, M., Finin, T., Glance, N. S., Nicolov, N., and Tseng, B. L., editors, *ICWSM*. The AAAI Press.

Gao, Q., Abel, F., Houben, G.-J., and Yu, Y. (2012). A comparative study of users' microblogging behavior on sina weibo and twitter. In Masthoff, J., Mobasher, B., Desmarais, M., and Nkambou, R., editors, *User Modeling, Adaptation, and Personalization*, volume 7379 of *Lecture Notes in Computer Science*, pages 88–101. Springer Berlin Heidelberg.

Gayo-Avello, D., Metaxas, P., and Mustafaraj, E. (2011). Limits of electoral predictions using twitter.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 42–47, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M., and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457:1012–1014. doi:10.1038/nature07634.

Glance, N. S., Hurst, M., and Tomokiyo, T. (2004). Blogpulse: Automated trend discovery for weblogs. In *IN WWW 2004 WORKSHOP ON THE WEBLOGGING ECOSYSTEM: AGGREGATION, ANALYSIS AND DYNAMICS*. ACM.

Godbole, N., Srinivasaiah, M., and Skiena, S. (2007). Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Goorha, S. and Ungar, L. (2010). Discovery of significant emerging trends. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 57–64, New York, NY, USA. ACM.

Grossman, L. (2009). Iran protests: Twitter, the medium of the movement. *Time Magazine*, 17.

Gupta, M., Gao, J., Zhai, C., and Han, J. (2012). Predicting future popularity trend of events in microblogging platforms. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10.

Han, J. (2005). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Hayes, A. S., Singer, J. B., and Ceppos, J. (2007). Shifting roles, enduring values: The credible journalist in a digital age. *Journal of Mass Media Ethics*, 22(4):262–279.

Hermida, A. (2009). The blogblog bbc: Journalism blogs at "the world's most trusted news organisation". *Journalism Practice*, 3(3):268–284.

Hermida, A. (2010). Twittering the news: The emergence of ambient journalism. *Journalism Practice*, 4(3):297–308.

Hoffman, M. D., Blei, D. M., and Bach, F. R. (2010). Online learning for latent dirichlet allocation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *NIPS*, pages 856–864. Curran Associates, Inc.

Honey, C. and Herring, S. (2009). Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, pages 1–10.

Hong, L. and Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA. ACM.

Huberman, B., Romero, D., and Wu, F. (2009). Social networks that matter: Twitter under the microscope. *First Monday*, 14(1):8.

Humphreys, L., Gill, P., Krishnamurthy, B., and Newbury, E. (2013). Historicizing new media: A content analysis of twitter. *Journal of Communication*, 63(3):413–431.

Ienco, D., Bonchi, F., and Castillo, C. (2010). The meme ranking problem: Maximizing microblogging virality. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 328–335.

Ihler, A., Hutchins, J., and Smyth, P. (2006). Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 207–216, New York, NY, USA. ACM.

Inouye, D. and Kalita, J. (2011). Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 298–306.

Jansen, B. J., Zhang, M., Sobel, K., and Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11):2169–2188.

Java, A., Song, X., Finin, T., and Tseng, B. (2007). Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM.

Kaplan, A. M. and Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2):105 – 113.

Kawai, Y., Kumamoto, T., and Tanaka, K. (2007). Fair news reader: Recommending news articles with different sentiments based on user preference. In Apolloni, B., Howlett, R., and Jain, L., editors, *Knowledge-Based Intelligent Information and Engineering Systems*, volume 4692 of *Lecture Notes in Computer Science*, pages 612–622. Springer Berlin Heidelberg.

Kim, H. D., Park, D. H., Lu, Y., and Zhai, C. (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10.

Kim, S. and Hovy, E. (2006). Identifying and analyzing judgment opinions. In *Proceedings of HLT/NAACL-2006*, pages 200–207.

Kleinberg, J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397.

Kontostathis, A., Galitsky, L., Pottenger, W., Roy, S., and Phelps, D. (2004). A survey of emerging trend detection in textual data mining. In Berry, M., editor, *Survey of Text Mining*, pages 185–224. Springer New York.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Lau, C. H., Li, Y., and Tjondronegoro, D. (2011). Microblog retrieval using topical features. In Voorhees, E. M. and Buckland, L. P., editors, *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011,*, Gaithersburg, Maryland. National Institute of Standards and Technology (NIST).

Lau, C. H., Tao, X., Tjondronegoro, D., and Li, Y. (2012). Retrieving information from microblog using pattern mining and relevance feedback. In *Data and Knowledge Engineering*, volume 7696 of *Lecture Notes in Computer Science*, pages 152–160. Springer Berlin Heidelberg.

Lau, C. H. and Tjondronegoro, D. (2010). Text mining in microblogs for real time topic and event monitoring. In *Super Computing (SC'10) Early Adopters PhD workshop*, New Orleans, USA.

Lee, C.-H., Wu, C.-H., and Chien, T.-F. (2011a). Burst: A dynamic term weighting scheme for mining microblogging messages. In Liu, D., Zhang, H., Polycarpou, M., Alippi, C., and He, H., editors, *Advances in Neural Networks – ISNN 2011*, volume 6677 of *Lecture Notes in Computer Science*, pages 548–557. Springer Berlin Heidelberg.

Lee, C.-H., Wu, C.-H., Yang, H.-C., and Wen, W.-S. (2012). Computing event relatedness based on a novel evaluation of social-media streams. In J. (Jong Hyuk) Park, J., Leung, V. C., Wang, C.-L., and Shon, T., editors, *Future Information Technology, Application, and Service*, volume 164 of *Lecture Notes in Electrical Engineering*, pages 697–707. Springer Netherlands.

Lee, C.-H., Yang, H.-C., Chien, T.-F., and Wen, W.-S. (2011b). A novel approach for event detection by mining spatio-temporal information on microblogs. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 254–259.

Lee, R., Wakamiya, S., and Sumiya, K. (2011c). Discovery of unusual regional social activities using geo-tagged microblogs. *World Wide Web*, 14(4):321–349.

Li, C., Sun, A., and Datta, A. (2012). Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 155–164, New York, NY, USA. ACM.

Li, J. and Rao, H. (2010). Twitter as a rapid response news service: An exploration in the context of the 2008 china earthquake. *The Electronic Journal of Information Systems in Developing Countries*, 42.

Li, Y., Algarni, A., and Zhong, N. (2010). Mining positive and negative patterns for relevance feature discovery. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 753–762, New York, NY, USA. ACM.

Liao, Y., Moshtaghi, M., Han, B., Karunasekera, S., Kotagiri, R., Baldwin, T., Harwood, A., and Pattison, P. (2012). Mining micro-blogs: Opportunities and challenges. In Abraham, A., editor, *Computational Social Networks*, pages 129–159. Springer London.

Lievrouw, L. A. (2005). New media design and development: Diffusion of innovations v social shaping of technology. *Handbook of New Media: Student Edition*, page 246.

Lin, C., Lin, C., Li, J., Wang, D., Chen, Y., and Li, T. (2012). Generating event storylines from microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 175–184, New York, NY, USA. ACM.

Liu, B. (2008). Opinion mining. Invited contribution to Encyclopedia of Database Systems.

Long, R., Wang, H., Chen, Y., Jin, O., and Yu, Y. (2011). Towards effective event detection, tracking and summarization on microblog data. In Wang, H., Li, S., Oyama, S., Hu, X., and Qian, T., editors, *Web-Age Information Management*, volume 6897 of *Lecture Notes in Computer Science*, pages 652–663. Springer Berlin Heidelberg.

Lu, R., Xu, Z., Zhang, Y., and Yang, Q. (2012). Life activity modeling of news event on twitter using energy function. In Tan, P.-N., Chawla, S., Ho, C., and Bailey, J., editors, *Advances in Knowledge Discovery and Data Mining*, volume 7302 of *Lecture Notes in Computer Science*, pages 73–84. Springer Berlin Heidelberg.

Magdy, W. (2013). Tweetmogaz: a news portal of tweets. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 1095–1096, New York, NY, USA. ACM.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, 1 edition.

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., and Miller, R. C. (2011). Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 227–236, New York, NY, USA. ACM.

Massoudi, K., Tsagkias, M., de Rijke, M., and Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. *Advances in Information Retrieval*, pages 362–367.

Mathioudakis, M. and Koudas, N. (2010). Twittermonitor: trend detection over the twitter stream. In *SIGMOD '10: Proceedings of the 2010 international conference on Management of data*, pages 1155–1158, New York, NY, USA. ACM.

Metzler, D., Cai, C., and Hovy, E. (2012). Structured event retrieval over microblog archives. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 646–655, Stroudsburg, PA, USA. Association for Computational Linguistics.

Metzler, D. and Croft, W. B. (2004). Combining the language model and inference network approaches to retrieval. *Inf. Process. Manage.*, 40:735–750.

Metzler, D., Dumais, S., and Meek, C. (2007). Similarity measures for short segments of text. In Amati, G., Carpineto, C., and Romano, G., editors, *Advances in Information Retrieval*, volume 4425 of *Lecture Notes in Computer Science*, pages 16–27. Springer Berlin Heidelberg.

Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B., and Magoulas, R. (2008). Twitter and the micro-messaging revolution: Communication, connections, and immediacy—140 characters at a time. An O'Reilly Radar Report . 54 pages.

Mishne, G. (2007). Using blog properties to improve retrieval. In *ICWSM*.

Mishne, G., de Rijke, M., Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., and Yavlinsky, A. (2006). A study of blog search. In *LNCS*, volume 3936, pages 289–301. Springer, Springer.

Naaman, M., Becker, H., and Gravano, L. (2011). Hip and trendy: Characterizing emerging trends on twitter. *Journal of the American Society for Information Science and Technology*.

Nagarajan, M., Purohit, H., and Sheth, A. P. (2010). A qualitative examination of topical tweet and retweet practices. In Cohen, W. W. and Gosling, S., editors, *ICWSM*. The AAAI Press.

Nagmoti, R., Teredesai, A., and De Cock, M. (2010). Ranking approaches for microblog search. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 153–157. IEEE.

Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. (2011a). Bad news travel fast: A content-based analysis of interestingness on twitter. *Proceedings of Web Science Conference*.

Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. (2011b). Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 183–188, New York, NY, USA. ACM.

Nichols, J., Mahmud, J., and Drews, C. (2012). Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA. ACM.

Nielsen, F. Å. (2011). A new anew: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

O'Connor, B., Balasubramanyan, R., Routledge, B., and Smith, N. (2010a). From tweets to polls: Linking text sentiment to public opinion time series.

O'Connor, B., Krieger, M., and Ahn, D. (2010b). Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Ounis, I., Macdonald, C., Lin, J., and Soboroff, I. (2011). Overview of the trec-2011 microblog track. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*.

Ozdikis, O., Senkul, P., and Oguztuzun, H. (2012). Semantic expansion of hashtags for enhanced event detection in twitter. In *Proceedings of the 1st International Workshop on Online Social Systems*.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Palen, L., Vieweg, S., Liu, S. B., and Hughes, A. L. (2009). Crisis in a networked world: Features of computer-mediated communication in the april 16, 2007, virginia tech event. *Social Science Computer Review*, 27(4):467–480.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Papacharissi, Z. and de Fatima Oliveira, M. (2012). Affective news and networked publics: The rhythms of news storytelling on #egypt. *Journal of Communication*, 62(2):266–282.

Parr, B. (2011). Julian assange defends wikileaks on "60 minutes". http://mashable.com/2011/01/30/julian-assange-60-minutes-video. Last accessed on Nov 01, 2013.

Paul, M. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health.

Petrovic, S., S., Osborne, M., and Lavrenko, V. (2010). Streaming first story detection with application to twitter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 181–189, Stroudsburg, PA, USA. Association for Computational Linguistics.

Phelan, O., McCarthy, K., and Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 385–388, New York, NY, USA. ACM.

Phuvipadawat, S. and Murata, T. (2010). Breaking news detection and tracking in twitter. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 3, pages 120–123. IEEE.

Popescu, A.-M. and Pennacchiotti, M. (2010). Detecting controversial events from twitter. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1873–1876, New York, NY, USA. ACM.

Qu, Y., Huang, C., Zhang, P., and Zhang, J. (2011). Microblogging after a major disaster in china: a case study of the 2010 yushu earthquake. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, CSCW '11, pages 25–34, New York, NY, USA. ACM.

Quincey, E. and Kostkova, P. (2010). Early warning and outbreak detection using social networking websites: The potential of twitter. In Kostkova, P., editor, *Electronic Healthcare*, volume 27 of *Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*, pages 21–24. Springer Berlin Heidelberg.

Ramage, D., Dumais, S., and Liebling, D. (2010). Characterizing microblogs with topic models. In *International AAAI Conference on Weblogs and Social Media*. The AAAI Press.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494. AUAI Press.

Rowe, M. and Stankovic, M. (2011). Aligning tweets with events: Automation via semantics. *Semantic Web Journal*.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 851–860, New York, NY, USA. ACM. p851-sakaki.pdf.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *INFORMATION PROCESSING AND MANAGEMENT*, pages 513–523.

Sankaranarayanan, J., Samet, H., Teitler, B. E., Lieberman, M. D., and Sperling, J. (2009). Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, GIS '09, pages 42–51, New York, NY, USA. ACM.

Schulz, A., Ristoski, P., and Paulheim, H. (2013). I see a car crash: Real-time detection of small scale incidents in microblogs. In *Proceedings of Social Media and Linked Data for Emergency Response (SMILE) Workshop at ESWC 2013*.

Sehgal, V. and Song, C. (2007/). Sops: Stock prediction using web sentiment. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, pages 21–26.

Sekiguchi, Y., Kawashima, H., Okuda, H., and Oku, M. (2006). Topic detection from blog documents using users interests. In *Mobile Data Management, 2006. MDM 2006. 7th International Conference on*, pages 108–108.

Shamma, D. A., Kennedy, L., and Churchill, E. F. (2009). Tweet the debates: understanding community annotation of uncollected sources. In *WSM '09: Proceedings of the first SIGMM workshop on Social media*, pages 3–10, New York, NY, USA. ACM.

Sharifi, B., Hutton, M.-A., and Kalita, J. (2010). Summarizing microblogs automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 685–688, Stroudsburg, PA, USA. Association for Computational Linguistics.

Shou, L., Wang, Z., Chen, K., and Chen, G. (2013). Sumblr: continuous summarization of evolving tweet streams. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '13, pages 533–542, New York, NY, USA. ACM.

Smith, D. (2011). Nelson mandela in hospital in johannesburg. http://www.theguardian.com/world/2011/jan/27/nelson-mandela-hospital-johannesburg. Last accessed on Nov 01, 2013.

Soboroff, I., McCullough, D., Lin, J., Macdonald, C., Ounis, I., and McCreadie, R. M. C. (2012). Evaluating real-time search over tweets. In Breslin, J. G., Ellison, N. B., Shanahan, J. G., and Tufekci, Z., editors, *ICWSM*. The AAAI Press.

Soboroff, I. and Robertson, S. (2003). Building a filtering test collection for trec 2002. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pages 243–250, New York, NY, USA. ACM.

Starbird, K. and Palen, L. (2012). (how) will the revolution be retweeted?: information diffusion and the 2011 egyptian uprising. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, pages 7–16, New York, NY, USA. ACM.

Stassen, W. (2011). Your news in 140 characters: exploring the role of social media in journalism. *Global Media Journal African Edition*, 4(1).

Steensen, S. (2011). Online journalism and the promises of new technology. *Journalism Studies*, 12(3):311–327.

Suh, B., Hong, L., Convertino, G., Chi, H., and Bernstein, M. (2010). Sensemaking with tweeting: Exploiting microblogging for knowledge workers. In *CHI 2010 Workshop on Microblogging*.

Tao, X., Zhou, X., Lau, C. H., and Li, Y. (2013). Personalised information gathering and recommender systems: techniques and trends. *ICST Transactions on Scalable Information Systems*, 13(1-3).

Teevan, J., Ramage, D., and Morris, M. (2011). #twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM.

Tjondronegoro, D., Tao, X., Sasongko, J., and Lau, C. H. (2011). Multi-modal summarization of key events and top players in sports tournament videos. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 471–478.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2011). Election forecasts with twitter: How 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):402–418.

Tzeng, Y.-S., Jiang, J.-Y., and Cheng, P.-J. (2012). Event duration detection on microblogging. In *Proceedings of the The 2012 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '12, pages 16–23, Washington, DC, USA. IEEE Computer Society.

USA Today (2011). 2011 nba all-star game roster. http://usatoday30.usatoday.com/sports/basketball/nba/2011-nba-all-star-roster.htm. Last accessed on Nov 01, 2013.

Uysal, I. and Croft, W. B. (2011). User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, CIKM '11, pages 2261–2264, New York, NY, USA. ACM.

Vieweg, S., Hughes, A. L., Starbird, K., and Palen, L. (2010). Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1079–1088, New York, NY, USA. ACM.

Vis, F. (2013). Twitter as a reporting tool for breaking news. *Digital Journalism*, 1(1):27–47.

Voorhees, E. M. and Buckland, L. P., editors (2011). *Proceedings of The Twentieth Text REtrieval Conference, TREC 2011, Gaithersburg, Maryland, November 15-18, 2011*. National Institute of Standards and Technology (NIST).

Weng, J. and Lee, B.-S. (2011). Event detection in twitter. In Adamic, L. A., Baeza-Yates, R. A., and Counts, S., editors, *ICWSM*. The AAAI Press.

Weng, J., Lim, E., Jiang, J., and He, Q. (2010a). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM.

Weng, J., Lim, E.-P., He, Q., and Leung, C.-K. (2010b). What do people want in microblogs? measuring interestingness of hashtags in twitter. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 1121–1126.

Wilkinson, D. and Thelwall, M. (2012). Trending twitter topics in english: An international comparison. *Journal of the American Society for Information Science and Technology*, 63(8):1631–1646.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005a). Opinionfinder: a system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005b). Recognizing contextual polarity in phrase-level sentiment analysis. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354, Morristown, NJ, USA. Association for Computational Linguistics.

Wu, S.-T., Li, Y., and Xu, Y. (2006). Deploying approaches for pattern refinement in text mining. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 1157–1161, Washington, DC, USA. IEEE Computer Society.

Wu, S.-T., Li, Y., Xu, Y., Pham, B., and Chen, P. (2004). Automatic pattern-taxonomy extraction for web mining. In *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on*, pages 242–248.

Xiao, F., Noro, T., and Tokuda, T. (2012). News-topic oriented hashtag recommendation in twitter based on characteristic co-occurrence word detection. In Brambilla, M., Tokuda, T., and Tolksdorf, R., editors, *Web Engineering*, volume 7387 of *Lecture Notes in Computer Science*, pages 16–30. Springer Berlin Heidelberg.

Xu, W., Feng, S., Wang, L., Wang, D., and Yu, G. (2012). Detecting hot topics in chinese microblog streams based on frequent patterns mining. In Wang, F., Lei, J., Gong, Z., and Luo, X., editors, *Web Information Systems and Mining*, volume 7529 of *Lecture Notes in Computer Science*, pages 637–644. Springer Berlin Heidelberg.

Yang, Y., Carbonell, J., Brown, R., Pierce, T., Archibald, B., and Liu, X. (1999). Learning approaches for detecting and tracking news events. *Intelligent Systems and their Applications, IEEE*, 14(4):32–43.

Yardi, S., Romero, D., Schoenebeck, G., and danah boyd (2009). Detecting spam in a twitter network. *First Monday*, 15(1).

Yarow, J. (2010). Twitter finally reveals all its secret stats.

Zhai, C. and Lafferty, J. (2001). Model-based feedback in the language modeling approach to information retrieval. In *Proceedings of the tenth international conference on Information and knowledge management*, CIKM '01, pages 403–410, New York, NY, USA. ACM.

Zhang, C. and Sun, J. (2012). Large scale microblog mining using distributed mb-lda. In *Proceedings of the 21st international conference companion on World Wide Web*, WWW '12 Companion, pages 1035–1042, New York, NY, USA. ACM.

Zhao, W., Jiang, J., Weng, J., He, J., Lim, E., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. *Advances in Information Retrieval*, pages 338–349.

Zhao, W. X., Jiang, J., He, J., Shan, D., Yan, H., and Li, X. (2010). Context modeling for ranking and tagging bursty features in text streams. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, CIKM '10, pages 1769–1772, New York, NY, USA. ACM.

Zhong, N., Li, Y., and Wu, S.-T. (2012). Effective pattern discovery for text mining. *Knowledge and Data Engineering, IEEE Transactions on*, 24(1):30 –44.